AI 연구의 발전과 사회적 책임: 기술에서 거버넌스까지

오상훈

2025년 3월부터 2026년 2월까지 한국과학기술정보연구원(KISTI)에서 객원연구원으로 활동하면서 최신 자료(보고서, 논문)들을 정리 및 축약하여 매월 마다 보고서를 작성하였다. 그 보고서들을 “기술(Technology)-응용(Application)-기대와 현실(Expectations and Reality)-위험(Risk)-통제(Governance & Control)”의 다섯 단계 프레임으로 정리하였다. 이를 통해 AI 연구가 성능 개선을 넘어 사회적 배포와 책임의 문제를 중시하게 되었음을 보여준다.

1단계. 기술 (Technology)

-지능을 만드는 방법에 대한 탐구-

핵심 내용

-Transformer 기반 모델의 구조적 이해와 확장 (NLP → Vision → Time Series → Energy-Based Transformers)

-SLM을 이용한 Agentic AI의 경제성 확보

-TextGrad를 통한 자연어 기반 최적화

-Class Imbalance 학습 기법

해당 조사

핵심 질문

-AI는 어떻게 학습하고, 어떻게 사고하는가?

의미

-구조·추론·학습 원리에 대한 이해

2단계. 응용 (Application)

-지능이 실제로 무엇을 할 수 있는가-

핵심 내용

-LLM을 활용한 과학 연구 자동화

-AI Co-Scientist, AlphaEvolve, AlphaGenome
-보건·생명과학·유전체·신약 개발

해당 조사

핵심 질문

-AI는 어디까지 인간의 지적 노동을 대체·보완할 수 있는가?

의미

-AI가 도구를 넘어 과학적 발견과 설계의 주체로 진입

3단계. 기대와 현실

-AI는 정말 세상을 바꾸고 있는가?-

핵심내용

-기술·응용 단계의 기대(hype)를 비판적으로 검토
-생산성, 사회 변화, 제도적 영향에서의 현실적 한계 분석

해당 조사

핵심 질문

-AI의 기술적 가능성과 실제 사회적 영향 사이에는 어떤 간극이 존재하는가?

의미

-위험(Risk)과 통제(Control) 논의가 왜 필요한지를 정당화하는 문제 제기 역할

4단계. 위험 (Risk)

-AI의 실패와 부작용-

핵심 내용

-LLM 환각(Hallucination)의 구조적 원인
-사회·경제·정보 생태계에 미치는 영향

해당 조사

핵심 질문

-AI는 언제, 왜, 어떻게 잘못되는가?

의미

-오류는 예외가 아니라 시스템적·통계적 결과임을 인식

5단계. 통제 (Governance & Control)

-지능을 안전하게 다루는 방법-

핵심 내용

-연쇄사고사슬 (COT, Chain-of-Thought) 모니터링

-모니터링 가능성(Monitorability) 평가 지표

-레드 팀 활동(Red Teaming) 기반 사전·사후 검증

-보건의료 분야의 윤리 및 거버넌스

해당 조사

핵심 질문

-AI를 신뢰 가능한 시스템으로 유지할 수 있는가?

의미

-가시성, 책임성, 사회적 합의의 중요성

트랜스포머: 자연어처리, 영상처리, 시계열 데이터로의 확장, 그리고 인간의 시스템 2 사고를 모델링한 에너지 기반 트랜스포머

LLM의 핵심 기술인 트랜스포머는 기존의 Seq2Seq 구조인 인코더-디코더를 따르면서도 RNN(Recurrent Neural Network)이나 CNN(Convolutional Neural Network)을 사용하지 않고 어텐션(Attention)으로 구현된 모델이다. 이 모델의 우수한 성능은 그 적용 영역을 자연어처리에서 넓혀나가, 영상에 적용된 ViT(Vision Tranformer)와 다변량 시계열 데이터에 적용된 TST(Time Series Transformer) 등이 있다. 여기서는 [A] 자연어처리의 표준으로 자리잡은 트랜스포머의 기본 원리를 살펴보고, [B] 영상처리 트랜스 포머(ViT)와 [C] 다변량 시계열 데이터 트랜스포머(TST)도 알아본다. 마지막으로 [D] 인간의 느리고 신중하며 분석적인 사고인 "시스템 2 사고"를 모방하여 모델의 학습과 사고 능력을 동시에 확장할 수 있는 새로운 패러다임으로 제시된 에너지 기반 프랜스포머(Energy-Based Tranformers) (2025.7.2.발표)에 대하여 알아본다.

조사문헌:

[A] A. Vaswani et al., "Attention is all you need," Proceedings of NIPS 2017.
- https://wikidocs.net/31379

[B] A. Dosovitskiy et al., "An image is worth 16x16 words: transformers for image recognition at scale," Proceedings of ICLR2021.
- I. Tolstikhin, et al.. "MLP-Mixer: An all-MLP Architecture for Vision," Proceedings of NeurIPS 2021, ArXiv:2105.01601v4

[C] G. Zerveas et al., "A Transformer-Based Framework for Multi-Variate Time Series Representation Learning," https://doi.org/10.48550/arXiv.2010.02803

[D] A. Gladstone et al., "Energy-Based Transformers are Scalable Learners and Thinkers," https://doi.org/10.48550/arXiv.2507.02092

트랜스포머: “Attention Is All You Need”

요약: 대표적인 시퀀스 변환(sequence transduction) 모델들은 일반적으로 인코더와 디코더를 포함하는 복잡한 순환(회귀) 신경망(RNN, Recurrent Neural Network)이나 합성곱 신경망(CNN, Convolutional Network Network)에 기반하고 있으며, 성능이 가장 뛰어난 모델들은 어텐션 메커니즘을 통해 인코더와 디코더를 연결한다. RNN이나 CNN 같은 복잡함을 완전히 없애고, 여기서는 오직 어텐션 메커니즘만을 기반으로 한 새로운 간단한 네트워크 아키텍처가 트랜스포머(Transformer)이다.

1 서론

RNN(Recurrent Neural Networks), LSTM(Long Short-Term Memory), 그리고 GRU (Gated Recurrent Unit Neural Networks) 등은 시퀀스 모델링과 변환(sequence modeling and transduction) 문제—예)언어 모델링(language modeling)과 기계 번역(machine translation)—에서 확고한 기술(state of the art)로 자리 잡아 왔으며, 지속적으로 많은 연구들이 순환 언어모델과 인코더-디코더 아키텍처의 한계를 넓혀왔다.

순환 모델은 일반적으로 입력 및 출력 시퀀스의 기호 위치(symbol positions)를 따라 연산을 분해한다. 이러한 위치들을 연산 시간의 단계들과 정렬시키며, 각 위치 t에 대해 이전 은닉 상태 h_{t-1}과 해당 위치의 입력을 함수로 하여 은닉 상태 h_t의 시퀀스를 생성한다. 이러한 순차적 특성 때문에, 학습에서의 병렬 처리가 불가능하며, 시퀀스 길이가 길어질수록 메모리 제약으로 인해 학습 예제 간 배치(batch) 처리도 제한을 받게 되어 문제가 심각해진다. 최근 연구들은 연산 분해 기법(factorization tricks)과 조건부 연산(conditional computation)을 통해 계산 효율성에서 상당한 향상을 이루었으며, 특히 조건부 연산은 모델 성능 또한 향상시켰지만, 이러한 발전에도 불구하고, 순차적 연산이라는 근본적인 제약은 여전히 남아 있다.

어텐션 메커니즘은 다양한 작업에서 뛰어난 시퀀스 모델링 및 변환(sequence modeling and transduction) 모델의 필수 요소로 자리 잡게 되었으며, 입력이나 출력 시퀀스 내에서의 거리와 무관하게 의존성을 모델링할 수 있게 해준다. 그러나 몇 가지 예외적인 경우를 제외하면, 이러한 어텐션 메커니즘은 RNN과 함께 사용되고 있다.

여기서는 순환구조를 배제하고, 어텐션 메커니즘만을 활용하여 입력과 출력 간의 전역적인 의존성을 학습하는 트랜스포머 모델 아키텍처를 제안한다. 트랜스포머는 훨씬 더 높은 수준의 병렬 처리를 가능하게 하여, 학습 시간을 줄여주면서도 우수한 성능을 달성해준다.

2 배경지식

순차연산을 줄이기 위한 노력: 트랜스포머에서는 이러한 연산 수를 상수(constant) 수준으로 줄였다. 다만, 이는 어텐션 가중치를 적용한 위치들의 평균(average)을 사용하는 특성으로 인해 효과적인 해상도가 줄어든다. 이 문제는 Multi-Head Attention 기법을 통해 보완한다.

셀프 어텐션은 하나의 시퀀스 내에서 서로 다른 위치들 간의 관계를 학습하여 해당 시퀀스의 표현을 계산해낸다. 셀프 어텐션은 독해, 추상적 요약(abstractive summarization), 텍스트 함의(textual entailment), 그리고 작업에 독립적인 문장 표현 학습(task-independent sentence representations) 등 다양한 작업에서 성공적으로 활용되어 왔다.

트랜스포머는 시퀀스에 정렬된 RNN이나 합성곱을 전혀 사용하지 않고, 입력과 출력의 표현을 계산하는 데 오직 셀프 어텐션만을 사용하는 최초의 변환(transduction) 모델이다.

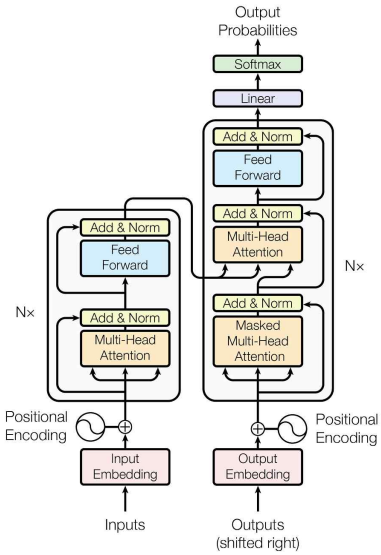


Figure 1: The Transformer - model architecture.

3 트랜스포머 모델 구조

대부분의 신경망 기반 시퀀스 변환 모델은 인코더-디코더 구조를 가지고 있다. 인코더는 입력 시퀀스 $(x_1,x_2,...,x_n)$를 받아, 연속적인 시퀀스 $\boldsymbol{z}=(z_1,z_2,...,z_n)$로 변환한다. 그런 다음, 디코더는 이 $\boldsymbol{z}$를 바탕으로 기호들로 이루어진 출력 시퀀스 $\boldsymbol{y}=(y_1,y_2,...,y_m)$을 한 번에 하나씩 생성한다. 각 단계에서 모델은 Autoregressive하게 작동하며, 이전에 생성된 기호들을 다음 기호를 생성할 때 추가 입력으로 사용한다.

트랜스포머는 이러한 전체적인 아키텍처를 따르며, 인코더와 디코더 모두에 대해 스택 형태의 셀프 어텐션과 포인트 와이즈(point-wise) 완전 연결 계층을 사용한다. 이는 그림 1의 왼쪽 절반(인코더)과 오른쪽 절반(디코더)에 각각 나타나 있다.

3.1 인코더 및 디코더

인코더: 인코더는 동일한 구조가 N층(N=6)으로 구성된 스택이다. 각 층은 두 개의 서브 레이어(sub-layer)로 이루어져 있다. 첫 번째 서브 레이어는 멀티-헤드 셀프 어텐션(MHSA, Multi-Head Self Attention) 메커니즘 층이며, 두 번째 서브 레이어는 위치별(position-wise) 완전 연결 전방향 신경회로망이다. 각 서브 레이어에는 잔차 연결(residual connection)이 적용되며, 그 뒤에 레이어 정규화(layer normalization)가 수행된다. 즉, 각 서브레이어의 출력은 LayerNorm(**x** + Sublayer(**x**))과 같이 계산된다. 여기서 Sublayer(**x**)는 해당 서브레이어가 수행하는 함수이다. 이러한 잔차 연결을 원활히 하기 위해, 모델의 모든 서브레이어와 임베딩(embedding) 레이어는 출력 차원을 $d_{\mathrm{model}}=512$로 고정한다.

디코더: 디코더 역시 동일한 구조의 N층(N=6)으로 구성된 스택이다. 각 인코더 층에 포함된 두 개의 서브 레이어 외에, 디코더는 세 번째 서브 레이어를 추가로 포함하며, 이는 인코더 스택의 출력에 대해 멀티-헤드 어텐션을 수행한다. 인코더와 마찬가지로, 각 서브 레이어에는 잔차 연결이 적용되고, 그 뒤에 레이어 정규화가 수행된다. 또한 디코더 스택의 셀프 어텐션 서브 레이어는 현재 위치가 이후 위치를 참조하지 못하도록 마스킹을 사용하였다. 이 마스킹은 출력 임베딩이 한 위치만큼 오른쪽으로 이동(offset)되는 것과 결합되어, 위치 i의 예측이 위치 i보다 작은 위치들의 출력에만 의존하도록 한다.

3.2 어텐션

어텐션 함수는 쿼리(query)와 키-밸류 쌍(key-value pairs)을 출력(output)으로 매핑한다. 이때 쿼리, 키, 밸류, 출력은 모두 벡터이다. 출력은 밸류들(values)의 가중치 합으로 계산되며, 각 밸류에 할당되는 가중치는 해당 쿼리와 해당 키 간의 호환(일치)성 함수(compatibility function)에 의해 결정된다.

3.2.1 스케일드 닷 프로덕트 어텐션 (Scaled Dot-Product Attention)

그림 2에서 보는 바와 같이 "스케일드 닷 프로덕트 어텐션(Scaled Dot-Product Attention)"

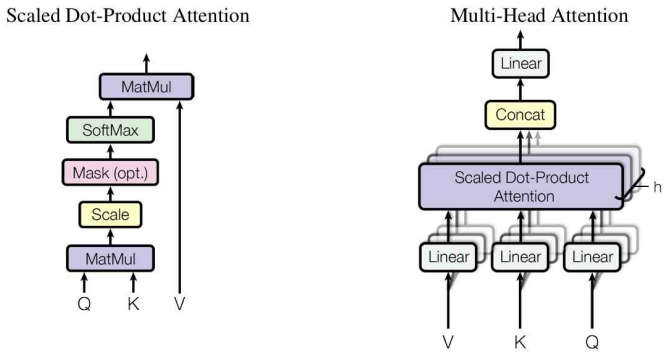


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

의 입력은 d_k차원의 쿼리와 키, 그리고 d_v차원의 밸류 벡터들로 구성된다. 여기서 쿼리와 모든 키 간의 내적(dot product)을 계산한 후, 각 값을 $\sqrt{d_k}$로 나누고, 그 결과에 소프트맥스(softmax) 함수를 적용하여, 밸류들에 대한 가중치(weights)를 얻게 된다.

실제로는 여러 개의 쿼리(query) 집합에 대해 어텐션 함수를 동시에 계산하며, 이 쿼리들을 하나의 행렬 Q로 묶어서 처리한다. 키와 밸류 역시 각각 행렬 K와 V로 구성된다. 그런 다음 출력 행렬은 다음과 같이 계산된다: $\mathrm{Attention(Q,K,V)=softmax(QK^T/\sqrt{d_k})V}$

가장 널리 사용되는 두 가지 어텐션 함수로 가산 어텐션과 내적(Dot-Product) 어텐션이 있다. 내적 어텐션은 본 방법과 동일하지만, 본 방법은 $1/\sqrt{d_k}$라는 스케일링 인자가 추가된 점이 다르다. d_k가 커질수록 내적의 값들이 커져 소프트맥스 함수가 매우 작은 기울기를 갖는 영역으로 밀려나는데, 이러한 효과를 완화하기 위해, 내적 결과를 $1/\sqrt{d_k}$로 스케일링한다.

3.2.2 멀티 헤드 어텐션(Multi-Head Attention)

쿼리, 키, 밸류를 각각 d_{model} 차원에서 단일 어텐션 함수로 처리하는 대신에, 쿼리, 키, 밸류에 대해 각각 서로 다른 "학습된 선형 변환(linear projection)"을 h번 수행하여, 각각을 d_k, d_k, d_v 차원으로 투영하는 방식이 더 효과적이다. 이렇게 투영된 쿼리, 키, 밸류 각각에 대해 병렬적으로 어텐션 함수를 수행하면, d_v 차원의 출력들이 생성된다. 이 출력들을 하나로 연결(concatenate)한 뒤, 다시 한 번 선형변환을 적용하여 최종 출력을 얻는다. 이 과정은 그림 2에 나타나 있다.

멀티-헤드 어텐션은 모델이 서로 다른 표현 서브 공간(representation subspaces)에서 서로 다른 위치의 정보에 동시에 주의를 기울일 수 있도록 한다. 반면, 단일 어텐션 헤드만 사용할 경우, 평균화(averaging)가 이러한 다양성을 억제한다.

$$MltiHead(Q,K,V))=Concat(head_1,...,head_h)W^O$$

여기서 $head_i=\mathrm{Attention(QW_i^Q,KW_i^K,VW_i^V)}$ 이고, 투영(projection)에 사용되는 행렬들은 다음과 같은 매개변수 행렬(parameter matrices)이다:

$$W_i^Q\in R^{d_{\mathrm{model}}\times d_k},\ W_i^K\in R^{d_{\mathrm{model}}\times d_k},\ W_i^V\in R^{d_{\mathrm{model}}\times d_v},\ W_i^O\in R^{hd_v\times d_{\mathrm{model}}}$$

여기서, $h=8, d_k=d_v=d_{\mathrm{model}}/h=64$를 사용하였다. 각 헤드의 차원이 줄어들었기 때문에, 전체 계산 비용은 전체 차원(d_{model})을 사용하는 단일 헤드 어텐션과 비슷하다.

3.2.3 어텐션 적용

트랜스포머에서 멀티 헤드 어텐션을 다음과 같이 세 가지 방식으로 사용한다:

• **"인코더 셀프 어텐션" 층**: 인코더 셀프 어텐션 층에서는 쿼리, 키, 밸류가 모두 같은 위치, 즉 이 경우에는 입력 혹은 인코더의 이전 층 출력에서 들어온다. 인코더 내에서 각 위치는 인코더의 이전 층에 있는 모든 위치에 주의(어텐션)를 기울일 수 있다.

• **"디코더 셀프 어텐션" 층(마스킹)**: 디코더의 셀프 어텐션 층은 디코더 내 각 위치가 해당 위치까지의 모든 위치에 주의를 기울일 수 있도록 한다. 오토레그레시브(auto-regressive) 특성을 유지하기 위해서는 왼쪽으로의 정보 흐름(leftward information flow)을 차단해야 한다. 이를 위해, 스케일드 닷 프로덕트 어텐션 내부에서 softmax의 입력 중 허용되지 않는 연결에 해당하는 값들을 마스킹($-\infty$로 설정)하여 구현한다.

• **"인코더-디코더 어텐션(encoder-decoder attention)" 층**: 쿼리는 이전 디코더 층에서 오고, 키와 밸류는 인코더의 출력에서 온다. 이로 인해 디코더의 모든 위치가 입력 시퀀스의 모든 위치에 주의를 기울일 수 있게 된다. 이 구조는 시퀀스-투-시퀀스(Seq2Seq) 모델에서의 전형적인 인코더-디코더 어텐션 메커니즘을 모방한 것이다.

3.3 위치별 전방향 신경망(Position-wise Feed-Forward Networks)

인코더와 디코더의 각 레이어에는 있는 완전연결 신경망은 각 위치에 대해 개별적으로 동일하게 적용된다. 이 신경망은 두 개의 선형 변환(linear transformation) 사이에 ReLU 활성화 함수가 삽입된 구조이며, 다음 수식으로 표현 가능하다: $FFN(x) = \max(0, x W_1 + b_1) W_2 + b_2$

선형 변환은 층별로 다른 파라미터를 사용한다. 입력과 출력의 차원은 $d_{model} = 512$이고, 내부층(은닉층)은 $d_{ff} = 2048$이다.

3.4 임베딩과 소프트맥스

다른 시퀀스 변환 모델들과 마찬가지로, 여기서도 학습된 임베딩을 사용하여 입력 토큰과 출력 토큰을 d_{model} 차원의 벡터로 변환한다. 또한, 그림 1에서 보는 바와 같이 디코더의 출력에서 학습된 선형 변환(linear transformation)과 소프트맥스 함수를 사용하여 다음 토큰의 확률을 예측한다. 트랜스포머는 두 개의 임베딩 층과 softmax 이전의 선형 변환 사이에서 동일한 가중치 행렬을 공유한다. 임베딩 층에서는 이 가중치에 $\sqrt{d_{model}}$을 곱해준다.

3.5 위치 인코딩(Positional Encoding)

트랜스포머는 순환이나 합성곱을 포함하지 않기 때문에, 시퀀스 내 토큰의 순서를 활용할 수 있도록 모델에 어떤 형태로든 상대적 또는 절대적인 위치 정보를 주어야 한다. 이를 위해, "위치 인코딩(positional encoding)"을 인코더와 디코더 스택의 맨 아래에서 입력 임베딩에 더해준다. 위치 인코딩은 임베딩과 같은 d_{model} 차원을 가지므로, 두 벡터를 더할 수 있다. 위치 인코딩에는 학습 가능한 방식과 고정된 방식 등 다양하게 선택할 수 있다. 여기서는 다음과 같이 서로 다른 주파수의 사인과 코사인 함수를 사용한다:

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}}), \quad PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}})$$

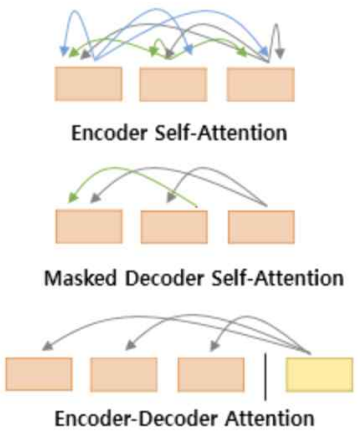
여기서, pos는 입력 문장에서 임베딩 벡터의 위치(position)를 나타내고, i는 임베딩 벡터 내의 차원 인덱스(index of dimension)이다. 임베딩 벡터 내의 각 차원의 인덱스가 짝수인 경우 사인 함수를 사용하고 홀수인 경우 코사인 함수를 사용한다. 사인파의 파장은 2π 부터 $10000 \cdot 2\pi$ 까지 기하급수적 증가(geometric progression) 형태를 따른다. 이 함수를 선택한 이유는, 모델이 상대적인 위치에 따라 쉽게 주의를 기울일 수 있도록 가정했기 때문이다. 왜냐하면, 어떤 고정된 오프셋 k에 대해서도, PE_{pos+k}는 PE_{pos}의 선형 함수로 표현될 수 있기

때문이다.

트랜스포머 내용 보완

"Attention Is All You Need" 논문에 나타나지 않지만 트랜스포머를 이해하기 위해 추가로 필요한 내용을 다음과 같이 정리하였다.
(https://wikidocs.net/31379)

어텐션층



첫번째 그림인 셀프 어텐션은 인코더에서 이루어지지만, 두번째 그림인 셀프 어텐션과 세번째 그림인 인코더-디코더 어텐션은 디코더에서 이루어진다. 셀프 어텐션은 본질적으로 쿼리, 키, 밸류가 동일한 경우를 말한다. 반면, 세번째 그림 인코더-디코더 어텐션에서는 쿼리가 디코더의 벡터인 반면에 키와 밸류는 인코더의 벡터이므로 셀프 어텐션이라고 부르지 않는다.

주의할 점은 여기서 쿼리, 키 등이 같다는 것은 벡터의 값이 같다는 것이 아니라 벡터의 출처가 같다는 뜻이다.
정리하면 다음과 같다.

인코더의 셀프 어텐션 : Query = Key = Value
디코더의 마스크드 셀프 어텐션 : Query = Key = Value
디코더의 인코더-디코더 어텐션 : Query → 디코더 벡터 / Key = Value → 인코더 벡터

Q, K, V 벡터 얻기

셀프 어텐션은 입력 문장의 단어 벡터들을 가지고 수행한다고 하였는데, 사실 셀프 어텐션은 인코더의 초기 입력인 d_{model} 차원을 가지는 단어 벡터들을 사용하여 셀프 어텐션을 수행하는 것이 아니라 우선 각 단어 벡터들로부터 Q벡터, K벡터, V벡터를 얻는 작업을 거친다. 이때 이 Q벡터, K벡터, V벡터들은 초기 입력인 d_{model} 차원 단어 벡터들보다 더 작은 차원을 가지는데, 논문에서는 $d_{model} = 512$ 차원을 가졌던 각 단어 벡터들을 64의 차원을 가지는 Q벡터, K벡터, V벡터로 $W_i^Q \in R^{d_{model} \times d_k}$, $W_i^K \in R^{d_{model} \times d_k}$, $W_i^V \in R^{d_{model} \times d_v}$ 를 이용하여 변환하였다. 64라는 값은 $h = 8, d_k = d_v = d_{model}/h = 64$에 의해 결정되었다.

스케일드 닷-프로덕트 어텐션(Scaled dot-product Attention)

Q, K, V 벡터를 얻었다면 지금부터는 기존 어텐션 메커니즘과 동일하다. 각 Q 벡터는 모든 K 벡터에 대해서 어텐션 스코어를 구하고, 어텐션 분포를 구한 뒤에 이를 사용하여 모든 V벡터를 가중치합을 하여 어텐션 값 또는 컨텍스트 벡터를 구하게 된다. 그리고 이를 모든 Q 벡터에 대해서 반복한다.

트랜스포머에서는 어텐션 챕터에 사용했던 내적만을 사용하는 어텐션 함수가 아니라 $1/\sqrt{d_k}$로 스케일링한 어텐션 함수를 사용한다. 이러한 함수를 사용하는 어텐션을 닷-프로덕트 어텐션(dot-product attention)에서 값을 스케일링하는 것을 추가하였다고 하여 스케일드 닷-프로덕트 어텐션(Scaled dot-product Attention)이라고 한다.

행렬 연산으로 일괄 처리하기

사실 각 단어에 대한 Q, K, V 벡터를 구하고 스케일드 닷-프로덕트 어텐션을 수행하였던 과정들은 벡터 연산이 아니라 행렬 연산을 사용하면 일괄 계산이 가능하다. 벡터 연산으로 설명하였던 이유는 이해를 돕기 위한 과정이고, 실제로는 행렬 연산으로 구현된다. 위의 과정을 벡터가 아닌 행렬 연산으로 알아보자. 우선, 각 단어 벡터마다 일일이 가중치 행렬을 곱하는 것이 아니라 문장 행렬에 가중치 행렬을 곱하여 Q 행렬, K 행렬, V 행렬을 구한다. 그런 다음 논문에 나온 다음 수식과 같이 계산된다: $\mathrm{Attention(Q,K,V)=softmax}\left(\mathrm{QK^T/\sqrt{d_k}}\right)\mathrm{V}$

멀티 헤드 어텐션(Multi-head Attention)

앞의 어텐션에서 $d_{\mathrm{model}}=512$ 차원을 가진 단어 벡터를 $h=8$로 나눈 64 차원을 가지는 Q, K, V 벡터로 바꾸고 어텐션을 수행하였다. 이제 h의 의미와 왜 d_{model} 차원을 가진 단어 벡터로 어텐션을 하지 않고 차원을 축소시킨 벡터로 어텐션을 수행하였는지 알아보자.

트랜스포머 연구진은 한 번의 어텐션을 하는 것보다 여러 번의 어텐션을 병렬로 사용하는 것이 더 효과적이라고 판단하였다. 그래서 d_{model} 차원을 h개로 나누어 d_{model}/h 차원을 가지는 Q, K, V에 대해서 h개의 병렬 어텐션을 수행한다. 다시 말해 위에서 설명한 어텐션이 h개로 병렬로 이루어지게 되는데, 이때 각각의 어텐션 값 행렬을 어텐션 헤드라고 부른다. 이때 가중치 행렬 W^Q, W^K, W^V의 값은 8개의 어텐션 헤드마다 전부 다르다. 어텐션을 병렬로 수행하는 것은 다른 시각으로 정보들을 수집하는 효과를 일으킨다.

병렬 어텐션을 모두 수행하였다면 모든 어텐션 헤드를 연결(concatenate)한다. 모두 연결된 어텐션 헤드 행렬의 크기는 $(SequenceLength, hd_v)=(SequenceLength, d_{\mathrm{model}})$이 된다.

어텐션 헤드를 모두 연결한 행렬은 또 다른 가중치 행렬 W^O를 곱하게 되는데, 이렇게 나온 결과 행렬이 멀티-헤드 어텐션의 최종 결과물이다. 이때 결과물인 멀티-헤드 어텐션 행렬은 인코더의 입력이었던 문장 행렬의 크기 $(SequenceLength, d_{\mathrm{model}})$와 동일하다. 다시 말해 인코더의 첫 번째 서브층인 멀티-헤드 어텐션 단계를 끝마쳤을 때, 인코더의 입력으로 들어왔던 행렬의 크기가 아직 유지되고 있다. 첫 번째 서브층인 멀티-헤드 어텐션과 두 번째 서브층인 위치별(position-wise) 전방향 신경망을 지나면서 인코더의 입력으로 들어올 때의 행렬의 크기는 계속 유지되어야 한다. 트랜스포머는 동일한 구조의 인코더를 쌓은 구조이다.

디코더의 끝단

디코더의 끝단에는 다중 클래스 분류 문제를 풀 수 있도록 vocab_size 만큼의 노드를 가지는 출력층을 추가한다.

...자연어 처리 교재의 트랜스포머 내용을 추가하자..

비젼 트랜스포머(ViT, Vision Transformer): “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”

요약: 트랜스포머는 자연어처리에서 사실상의 표준으로 자리 잡았다. 한편, 컴퓨터 비전 분야에서는 어텐션이 CNN과 함께 사용되거나, CNN의 일부 구성 요소를 대체하는 방식으로 적용되며, 전체 구조는 그대로 유지하는 경우가 많다. 그러나, 이러한 CNN에 대한 의존 없이, 여기에서 제안하는 ViT(Vision Transformer)는 이미지 패치들의 시퀀스에 순수한 트랜스포머(pure transformer)를 직접 적용해도 이미지 분류 작업에서 매우 좋은 성능을 낼 수 있다. 대규모 데이터로 사전 학습(pre-training)을 거친 후, 중간 규모 또는 소규모 이미지 인식 벤치마크에 전이 학습(transfer learning)을 수행하면, 최신 CNN들과 비교해도 뛰어난 결과를 얻을 수 있다. 그리고, 학습시키는 데 필요한 계산 자원도 훨씬 적게 든다.

1 서론

셀프 어텐션 기반 아키텍처인 트랜스포머는 자연어처리(NLP)에서 최선호 모델이며, 대규모 텍스트 코퍼스로 사전 학습(pre-train)을 수행한 후, 작은 크기의 태스크 특화 데이터 셋으로 파인 튜닝(fine-tune)하는 방식으로 사용되고 있다. 트랜스포머의 계산 효율성과 확장성 덕분에, 1000억 개가 넘는 파라미터를 가진 초대형 모델도 학습하는 것이 가능해졌다.

컴퓨터 비전 분야는 CNN 기반 아키텍처가 주류를 이루고 있는 상황에서 NLP 분야의 성공에서 영감을 받아, 여러 연구에서는 CNN 유사 아키텍처에 셀프 어텐션을 결합하려는 시도를 하고 있으며, 일부 연구는 합성곱을 완전히 대체하려 하였다.

여기서는 NLP에서 트랜스포머의 성공 사례에 착안하여 표준 트랜스포머를 가능한 최소한의 수정만으로 이미지에 직접 적용하였다. 이를 위해, 이미지를 작은 패치들로 나눈 뒤, 이 패치들의 선형 임베딩 시퀀스를 트랜스포머에 입력한다. 이때 이미지 패치들은 NLP에서의 토큰(단어)처럼 동일하게 취급된다.

ViT는 CNN이 본질적으로 가지고 있는 귀납적 편향(inductive bias)(예: 평행 이동 등가성이나 국소성)이 부족하기 때문에, 충분한 양의 데이터 없이 학습될 경우 일반화 성능이 떨어질 수 있다. 이는 ImageNet과 같은 중간 규모의 데이터셋에서 학습할 경우, ResNet보다 몇 퍼센트 낮은 수준의 정확도만을 보여줌으로 확인할 수 있다.

하지만 ViT를 더 큰 데이터셋(1,400만~3억 개 이미지)으로 학습시키면 귀납적 편향을 뛰어넘는 효과를 보인다. 즉, ViT는 충분히 큰 규모로 사전 학습(pre-training)되고, 데이터가 적은 작업(task)에 전이 학습되었을 때 우수한 성능을 보인다.

2. 관련 연구

이미지를 대상으로 셀프 어텐션을 단순하게 적용하면, 모든 픽셀이 다른 모든 픽셀에 주의를 기울여야 하므로, 계산 비용이 픽셀 수의 제곱에 비례하게 되어, 현실적이지 않다. 따라서 이미지 처리에서 트랜스포머를 적용하기 위해, 여러 가지 근사 방법(approximation)이 시도되었다. 예를 들어, Parmar et al.은 전역적인 어텐션이 아닌, 각 쿼리 픽셀에 대해 국소 영역(local neighborhood)에서만 셀프 어텐션을 적용했다. 이러한 국소 multi-head 내적 셀프 어텐션 블록은 합성곱(convolution)을 완전히 대체할 수 있다. 다른 접근으로는, Sparse Transformer가 있으며, 이는 전역 셀프 어텐션을 확장 가능하게 만들기 위한 근사 기법을

사용하여 이미지에 적용할 수 있도록 했다. 어텐션을 서로 다른 크기의 블록(block) 단위로 적용하거나 극단적인 경우에는 하나의 축(axis)에서만 어텐션을 적용하는 방법도 있다. 이러한 특화된 어텐션 아키텍처들은 컴퓨터 비전 작업에서 유망한 결과를 보여주지만, 하드웨어 가속기에서 효율적으로 구현하기 위해서는 복잡한 엔지니어링이 필요하다.

Cordonnier et al. 모델은 입력 이미지에서 2×2 크기의 패치를 추출한 뒤, 그 위에 전면적인 셀프 어텐션(full self-attention)을 적용하였다. 이 모델은 ViT와 매우 유사하지만, ViT는 대규모 사전 학습으로 일반적인(바닐라) 트랜스포머도 최첨단 CNN과 경쟁하거나 이를 능가할 수 있음을 보여준다는 점에서 더 나아간 결과로 볼 수 있다. 또한, Cordonnier et al. 모델은 2×2 픽셀의 작은 패치 크기만을 사용하기 때문에 모델이 저해상도 이미지에만 적용 가능한 반면, ViT는 중간 해상도 이미지까지도 처리할 수 있다.

3 비전 트랜스포머

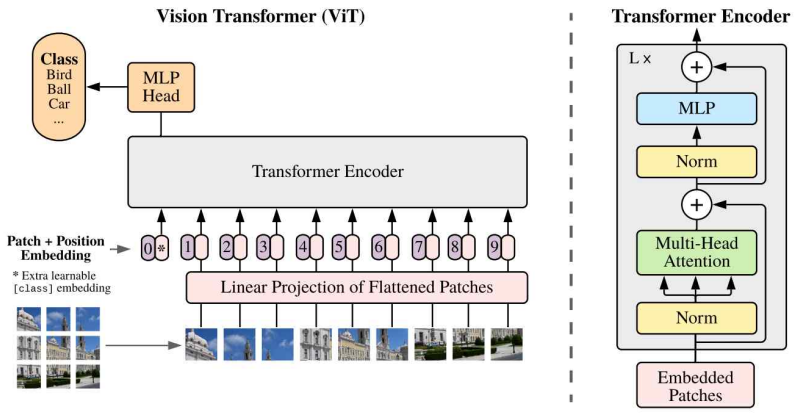


Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

모델의 개요는 그림 1에 나타나 있다. 트랜스포머는 1차원 토큰 임베딩 시퀀스를 입력으로 받으므로, 2차원 이미지를 1차원 토큰 형태로 처리하여야 한다. 이를 위해, 이미지 $\boldsymbol{x} \in R^{H \times W \times C}$를 평탄화된 2차원 패치들의 시퀀스 $\boldsymbol{x}_p \in R^{N \times (P^2 C)}$로 변환한다. 여기서 (H, W)는 원본 이미지의 해상도이고, C는 채널 수, (P, P)는 각 이미지 패치의 해상도이다. $N = HW/P^2$는 생성되는 패치의 수로, 이는 트랜스포머의 실질적인 입력 시퀀스 길이로 사용된다. 트랜스포머는 모든 계층에서 일정한 잠재 벡터 크기 D를 사용하므로, 패치들을 평탄화한 뒤 학습 가능한 선형 투영(식 1)을 통해 D차원으로 매핑한다. 이 투영의 출력을 패치 임베딩(patch embeddings)이라고 한다.

BERT(Bidirectional Encoder Representation from Tranformers)의 [class] 토큰과 유사하게, ViT는 임베딩된 패치 시퀀스 앞에 학습 가능한 임베딩($\boldsymbol{z}_0^0 = \boldsymbol{x}_{class}$)을 추가한다. 트랜스포머 인코더의 출력에서 이 인코딩의 상태($\boldsymbol{z}_L^0$)를 이미지 표현 $\boldsymbol{y}$(Eq. 4)로 사용한다. 사전 학습과 미세 조정(fine-tuning) 모두에서, 분류 헤드(classification head)가 $\boldsymbol{z}_L^0$에 연결된다. 분류 헤드는 사전 학습 시에는 하나의 은닉층을 가진 MLP로, 미세 조정 시에는 단일 선형 계층으로 구현된다.

$$\boldsymbol{z}_0 = [\boldsymbol{x}_{class};\ \boldsymbol{x}_p^1\boldsymbol{E}\ ;\boldsymbol{x}_p^2\boldsymbol{E}\ ;\ldots;\boldsymbol{x}_p^N\boldsymbol{E}] + \boldsymbol{E}_{pos},\ \boldsymbol{E} \in R^{(P^2C)\times D}, \boldsymbol{E}_{pos} \in R^{(N+1)\times D} \quad (1)$$

$$\boldsymbol{z}'_l = MSA(LN(\boldsymbol{z}_{l-1})) + \boldsymbol{z}_{l-1},\ l = 1,2,\ldots,L \quad (2)$$

$$\boldsymbol{z}_l = MLP(LN(\boldsymbol{z}'_l)) + \boldsymbol{z}'_l,\ l = 1,2,\ldots,L \quad (3)$$

$$\boldsymbol{y} = LN(\boldsymbol{z}_L^0) \quad (4)$$

패치 임베딩에 위치 임베딩을 추가하여 위치 정보를 보존한다. 2D 위치 임베딩을 사용해도 성능 향상이 크지 않아, 표준 학습 가능한 1차원 위치 임베딩을 채택하였다. 이렇게 생성된 임베딩 벡터 시퀀스는 인코더의 입력으로 사용된다.

* 학습 가능(learnable) 위치 임베딩: 임의의 값(주로 정규분포)으로 초기화되고, 학습 과정에서 데이터에 맞게 조정된다. 이는 고정된(예: 사인/코사인 함수 기반) 위치 임베딩과 다르다.

트랜스포머 인코더는 다중 헤드 셀프 어텐션(MSA) 블록(식 2)과 MLP 블록(식 3)이 번갈아 가며 쌓인 구조로 이루어져 있다. 각 블록 앞에는 Layernorm(LN)이 적용되며, 각 블록 뒤에는 잔차 연결이 추가된다. 이러한 과정을 식 (1)~(4)에 나타내었다.

전체 학습 구조

ViT에서 학습은 트랜스포머 인코더와 MLP(분류) 헤드의 모든 가중치가 동시에 업데이트되는 엔드-투-엔드(end-to-end) 방식으로 이루어진다. 즉, 트랜스포머 인코더와 MLP Head의 모든 가중치는 손실 함수의 값에 따라 역전파를 통해 동시에 업데이트된다. 이로 인해 모델 전체가 입력 이미지와 정답 레이블의 관계를 점차 학습한다.

1. 순전파(Forward Pass)

입력 이미지는 패치로 분할되어 임베딩되고, 위치 임베딩과 [class] 토큰이 추가된다. 이 시퀀스가 트랜스포머 인코더를 통과하며, 마지막 [class] 토큰 벡터가 이미지 전체를 대표하는 특징 벡터로 추출된다. 이 벡터가 MLP Head에 입력되어 최종 분류 결과를 출력한다.

2. 역전파(Backpropagation) 및 가중치 업데이트

예측 결과와 실제 정답 레이블을 비교하여 손실 함수를 계산하고, 계산된 손실을 기준으로 역전파가 수행된다. MLP Head의 가중치에서 시작해, 트랜스포머 인코더의 각 계층(멀티헤드 어텐션, MLP, LayerNorm 등)까지 모든 파라미터에 대해 그래디언트가 계산되고, 최적화 알고리즘이 각 파라미터(가중치, 바이어스 등)를 업데이트한다.

TST(Time Series Transformer): "A Transformer-Based Framework for Multivariate Time Series Representation Learning"

요약: 트랜스포머 기반 프레임워크를 이용한 다변량 시계열의 비지도 표현 학습(unsupervised representation learning)을 제안한다. 사전 학습된 모델은 회귀, 분류, 예측, 결측값 보간 등 다양한 다운스트림 작업에 잠재적으로 활용될 수 있다.

1 서론

다변량 시계열(MTS, Multivariate time series)은 과학, 의학, 금융, 공학, 산업 등 매우 다양한 분야에서 널리 사용되는 중요한 데이터 유형이다. MTS는 일반적으로 시간에 따라 동기화된 여러 변수(예: 서로 다른 물리량의 동시 측정값)의 변화를 나타내지만, 더 일반적으로는 하나의 공통된 독립 변수에 정렬된 여러 종속 변수들의 집합을 나타낼 수도 있다. 예를 들어, 빛의 주파수에 따른 다양한 조건에서의 흡수 스펙트럼 등이 이에 해당한다. 최근 "빅데이터" 시대가 도래하면서 MTS 데이터가 풍부해졌음에도 불구하고, 특히 라벨이 달린 데이터의 가용성은 훨씬 더 제한적이다. 방대한 데이터에 라벨을 부여하는 작업은 많은 시간과 노력, 특수한 인프라, 또는 도메인 전문 지식이 필요하기 때문에 비용이 많이 들거나 실질적으로 불가능한 경우가 많다. 이러한 이유로, 앞서 언급한 모든 분야에서는 소량의 라벨 데이터만 사용하거나, 기존에 존재하는 방대한 비라벨 데이터를 활용하여 높은 정확도를 제공할 수 있는 방법에 대한 관심이 매우 높다.

단변량 및 다변량 시계열을 위한 다양한 모델링 접근법이 있으며, 최근에는 딥러닝 모델이 예측, 회귀, 분류와 같은 작업에서 기존의 최신 기법에 도전하거나 이를 대체하고 있기도 하다. 그러나 컴퓨터 비전이나 자연어 처리와 같은 분야와 달리, 시계열 분야에서 딥러닝의 우위는 확고하지 않다.

본 연구에서는 최초로 트랜스포머 인코더를 다변량 시계열의 비지도 표현 학습 뿐만 아니라 시계열 회귀 및 분류 작업에도 적용하는 방안을 탐구한다. 트랜스포머는 처음에 기계번역을 위해 제시되었으며 이제는 사실상 모든 NLP 작업에서 최신 성능을 보여주고 있다. 트랜스포머가 NLP 분야에서 널리 성공할 수 있었던 주요 요인 중 하나는 비지도 사전학습을 통해 자연어를 표현하는 방법을 학습하는 능력을 지녔기 때문이다. 트랜스포머 모델은 멀티헤드 어텐션(multi-headed attention) 메커니즘을 기반으로 하며, 이 메커니즘은 여러 가지 중요한 장점을 제공하여 트랜스포머가 시계열 데이터에 특히 적합하다.

NLP 분야에서 트랜스포머 모델의 비지도 사전학습을 통해 얻은 인상적인 결과를 얻은 것에 착안하여, 본 연구에서는 비라벨 데이터를 활용할 수 있는 일반적으로 적용 가능한 방법론(프레임워크)을 개발하였다. 이 방법론은 입력 디노이징(오토리그레시브, autoregressive) 목적함수를 통해 트랜스포머 모델을 먼저 학습시켜, 다변량 시계열의 밀집 벡터 표현(dense vector representations)을 추출한다. 이렇게 사전 학습된 모델은 이후 회귀, 분류, 결측치 보간, 예측 등 다양한 다운스트림 작업에 적용될 수 있다.본 연구는 트랜스포머 모델이 매우 제한된 수의 학습 샘플(수백 개 수준)만으로도 모든 기존 SOTA 모델링 접근법을 설득력 있게 능가할 수 있음을 보였다. 비지도 학습이 추가적인 비라벨 데이터 샘플을 사용하지 않고도 다변량 시계열의 분류 및 회귀에서 지도 학습보다 우위를 보인 것은 이것이 처음이다. 또한, NLP 분야에서 최고의 트랜스포머 모델은 수십억 개의 파라미터를 가지고 다수의 GPU나

TPU에서 며칠에서 몇 주에 걸쳐 사전학습이 필요하다는 일반적인 인식과 달리, 본 연구의 모델들은 최대 수십만 개의 파라미터만을 사용하며, CPU에서도 실질적으로 학습이 가능하다.

2 관련 연구

본 연구는 트랜스포머의 활용을 특정 생성적 작업(전체 인코더-디코더 아키텍처가 필요한)에 국한하지 않고, 비지도 사전학습이 가능하며 약간의 수정만으로 다양한 다운스트림 작업에 쉽게 적용할 수 있는 프레임워크로 일반화하는 것을 목표로 한다. 이는 BERT가 번역 모델을 비지도 학습 기반의 범용 프레임워크로 전환하여, NLP 분야에서 트랜스포머의 지배적인 위치를 확립한 방식과 유사하다.

3 제안 방법(Time Series Transformer)

3.1 기반 모델

제안 방법의 핵심은 트랜스포머 인코더만을 사용하고 디코더 부분은 사용하지 않는 것이다. 공통적으로 사용되는 모델의 일반적인 부분의 개략도가 그림 1에 나와 있다. 여기서는 이 모델이 이산적인 단어 인덱스 시퀀스가 아닌 다변량 시계열 데이터와 호환되도록 한 변경 사항을 제안한다.

각 학습 샘플은 m개의 변수로 이루어진 길이가 w의 다변량 시계열이며, 이는 w개의 특성 벡터 $\boldsymbol{x}_t \in R^m$의 시퀀스를 구성한다: $\boldsymbol{X} \in R^{m \times w} = [\boldsymbol{x}_1, \boldsymbol{x}_{2,\ldots,} \boldsymbol{x}_w]$. 특성 벡터 $\boldsymbol{x}_t$가 먼저 정규화된 후, 트랜스포머 모델 시퀀스 요소 표현(일반적으로 모델 차원이라고 부름)의 차원인 d-차원 벡터 공간으로 선형 투영된다:

$$\boldsymbol{u}_t = \boldsymbol{W}_p \boldsymbol{x}_t + \boldsymbol{b}_p. \tag{1}$$

여기서, $\boldsymbol{W}_p \in R^{d \times m}$, $\boldsymbol{b}_p \in R^d$는 학습 가능한 파라미터이고, $\boldsymbol{u}_t \in R^d \ (t = 0, \ldots, w)$는 모델 입력 벡터이다. 이 벡터들은 NLP 트랜스포머의 단어 벡터에 해당한다. 이 벡터들은 위치 인코딩을 더하고 해당 행렬들과 곱해진 후, 셀프 어텐션 계층의 쿼리, 키, 밸류가 된다.

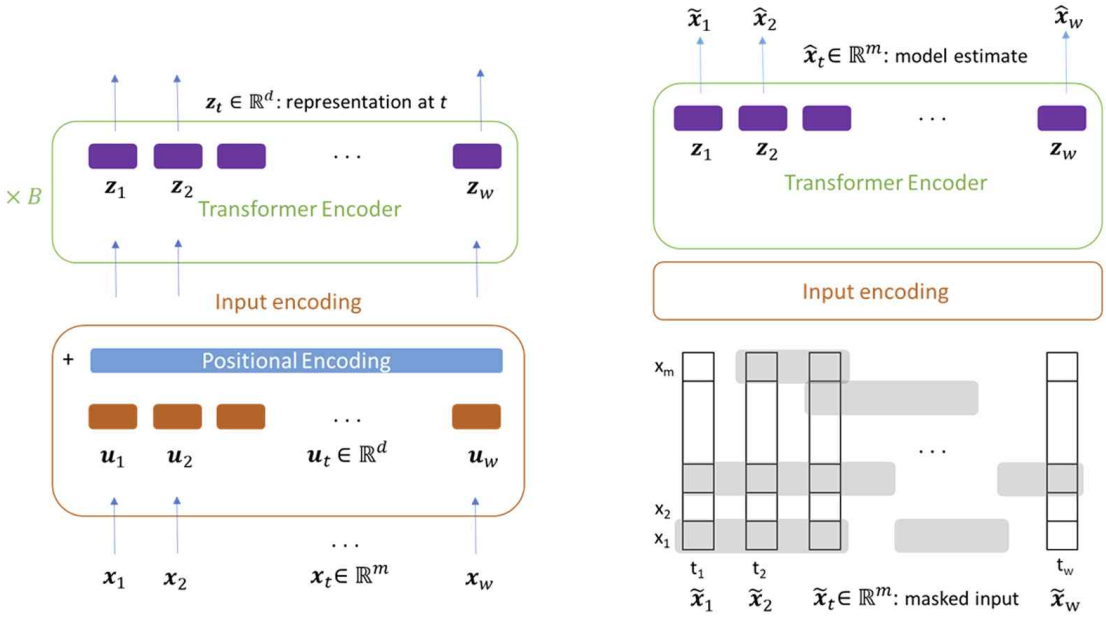


Figure 1: **Left:** Generic model architecture, common to all tasks. The feature vector $\mathbf{x_t}$ at each time step t is linearly projected to a vector $\mathbf{u_t}$ of the same dimensionality d as the internal representation vectors of the model and is fed to the first self-attention layer to form the keys, queries and values after adding a positional encoding. **Right:** Training setup of the unsupervised pre-training task. We mask a proportion r of each variable sequence in the input independently, such that across each variable, time segments of mean length l_m are masked, each followed by an unmasked segment of mean length $l_u = \frac{1-r}{r} l_m$. Using a linear layer on top of the final vector representations $\mathbf{z_t}$, at each time step the model tries to predict the full, uncorrupted input vectors $\mathbf{x_t}$; however, only the predictions on the masked values are considered in the Mean Squared Error loss.

마지막으로, 트랜스포머는 입력의 순서에 민감하지 않은 전방향 구조이기 때문에, 시계열의 순차적인 특성을 인식할 수 있도록 위치 인코딩 $W_{pos} \in R^{d \times w}$를 입력 벡터 $U \in R^{d \times w} = [u_1, u_2, ..., u_w]$에 더한다: $U' = U + W_{pos}$.

여기서는 완전히 학습 가능한 위치 인코딩을 사용한다. 이는 본 연구에서 제시된 모든 데이터셋에서 더 나은 성능을 보였기 때문이다. 모델의 성능을 바탕으로, 위치 인코딩이 시계열의 수치 정보에 일반적으로 큰 영향을 미치지 않는다는 점도 관찰할 수 있었으며, 이는 단어 임베딩의 경우와 유사하다. 그 이유가 위치 인코딩이 투영된 시계열 샘플이 존재하는 공간과는 다소 직교하는(거의 겹치지 않는) 다른 부분 공간을 차지하도록 학습되기 때문이라고 생각된다. 이러한 근사적인 직교성 조건은 고차원 공간에서 훨씬 더 쉽게 만족될 수 있다.

시계열 데이터와 관련하여 중요한 고려 사항 중 하나는 개별 샘플의 길이가 상당히 다양할 수 있다는 것이다. 이 문제는 제안한 프레임워크에서 효과적으로 처리된다. 전체 데이터셋에 대해 최대 시퀀스 길이 w를 설정한 후, 더 짧은 샘플은 임의의 값으로 패딩하고, 패딩된 위치에 대해 어텐션 점수에 큰 음수 값을 추가하는 패딩 마스크를 생성한다. 그런 다음 Softmax 함수를 사용해 셀프 어텐션 분포를 계산하기 전에 이 마스크를 적용한다. 이 방식은 모델이 패딩된 위치를 완전히 무시하게 하면서, 대규모 미니배치에서 샘플을 병렬로 처리할 수 있도록 해준다.

NLP에서 트랜스포머는 셀프 어텐션을 계산한 후와 각 인코더 블록의 피드 포워드 부분 이

후에 레이어 정규화(layer normalization)를 사용하며, 이는 Vaswani 등(2017)이 처음 제안한 것처럼 배치 정규화(batch normalization)보다 상당한 성능 향상을 가져왔다. 그렇지만, 본 연구에서는 대신 배치 정규화를 사용한다. 그 이유는 배치 정규화가 시계열 데이터에서 발생할 수 있는 이상치(outlier) 값의 영향을 완화할 수 있기 때문이며, 이러한 문제는 NLP의 단어 임베딩에서는 발생하지 않는다. 또한 NLP에서 배치 정규화의 낮은 성능은 주로 샘플 길이(즉, 대부분의 작업에서 문장)의 극단적인 변동성 때문으로 알려져 있다. 반면, 여기서 다루는 데이터셋에서는 이러한 변동성이 훨씬 적다.

“Energy-Based Transformers are Scalable Learners and Thinkers“

요약: 추론 시점(inference-time) 계산 기법은 인간의 시스템 2 사고(System 2 Thinking)처럼 모델 성능 향상을 위해 최근 각광받고 있으나, 대부분의 기존 접근법들은 몇 가지 한계점을 가지고 있다. 예를 들어, 특정 모달리티(예: 텍스트)에만 적용 가능하거나, 특정 문제(예: 수학이나 코딩과 같이 검증 가능한 도메인)에 한정되며, 또는 지도학습이나 추가 훈련(예: 검증기 또는 검증 가능한 보상 등)을 필요로 한다. 여기서는 “시스템 2 사고 방식을 일반화하여, 비지도 학습만으로 사고하는 법을 학습하는 모델을 개발할 수 있는가?”라는 질문을 제기한다. 이에 대한 답이 “그렇다”는 것을 발견했으며, 그 방법은 입력과 후보-예측 간의 호환성을 명시적으로 검증하는 법을 학습하고, 예측 문제를 이 검증기를 활용한 최적화 문제로 재구성하는 것이다. 구체적으로, EBTs(Energy-Based Transformers) 라는 새로운 형태의 EBMs (Energy-Based Models)을 훈련시켰다. 이 모델은 입력과 후보-예측 쌍마다 에너지(정규화되지 않은 확률) 값을 할당하고, 에너지 최소화 기반의 경사하강법을 통해 수렴에 이를 때까지 예측을 수행한다. 이러한 구조는 비지도 학습으로부터 시스템 2 사고가 자연스럽게 나타나게 하며, 모달리티나 문제에 구애받지 않는 범용적인 접근을 가능하게 한다. EBT는 모델의 학습과 사고 능력을 동시에 확장할 수 있는 유망한 새로운 패러다임이라 할 수 있다.

1 서 론

심리학에서는 인간의 사고를 “시스템 1 사고”와 “시스템 2 사고”로 분류한다. 시스템 1 사고는 빠르고 직관적이며 자동적인 반응이 특징으로, 이전의 경험에 의존하여 단순하거나 익숙한 문제를 해결한다. 반면, 시스템 2 사고는 느리고 신중하며 분석적인 사고로, 복잡한 정보를 처리하기 위해 의식적인 노력과 논리적 추론을 필요로 한다. 시스템 2 사고는 수학, 프로그래밍, 다단계 추론, 또는 새로운 분포 밖(out-of-distribution) 상황과 같이 자동적인 패턴 인식을 넘어서는 복잡한 문제를 해결할 때 필수적이며, 정밀성과 깊이 있는 이해가 중요한 경우에 해당한다. 현재 인공지능 모델은 시스템 1 사고에 적합한 과제에서는 뛰어나지만, 시스템 2 능력이 요구되는 과제에서는 여전히 어려움을 겪고 있다. 그 결과, 시스템 2 사고 능력을 추구하는 연구는 AI 연구자들 사이에서 점점 더 큰 관심을 받게 되었다. 이러한 “추론(reasoning) 모델”들은 수학 및 코딩 벤치마크에서 특히 뛰어난 성능을 보이며, 이는 모델이 사고에 더 많은 시간을 할애하기에 가능해졌다. 그러나 특히 오픈소스로 공개된 R1 모델의 학습 방법에 대한 공개 정보를 살펴보면, 이러한 모델들이 사용하는 강화학습 기반의 학습 접근법은 규칙 기반 보상으로 정답을 쉽게 검증할 수 있는 수학이나 코딩과 같은 도메인에서만 효과적으로 작동함을 알 수 있다. 이러한 제한점은 적용 가능성을 소수의 문제 유형으로 좁히며, 결과적으로 글쓰기와 같은 다른 작업에서는 오히려 성능 저하를 초래하는 경우가 많다. 게다가 최근 연구에 따르면, 이 강화학습 기반 접근법은 새로운 추론 패턴을 유도하는 것이 아니라, 기존의 기본 모델이 이미 알고 있는 추론 패턴의 발생 확률만을 증가시킨다는 증거가 있다. 이는 탐색(exploration)이 요구되는 문제들에 대한 성능을 제한하는 요인으로 작용할 수 있다.

이와 유사하게, 확산 모델과 순환 신경망(RNN)에서도 시스템 2 사고를 구현하려는 시도가 있다. 확산 모델은 노이즈 제거 단계를 통한 반복적인 추론(inference)을 지원하며, 노이즈 제거 단계를 늘리면 성능이 향상된다. 그러나 일반적으로 학습 당시의 노이즈 제거 단계를 초과

하면 성능 향상이 제한되는 경향이 있다. 또한, 단순히 노이즈 제거 단계를 늘리는 것 외에도, 확산 모델이 시스템 2 사고 능력을 향상시키기 위해서는 외부 검증기의 도움이 필요하다. RNN 역시 순환 상태 업데이트를 통해 반복적 계산을 수행할 수 있지만, 대부분의 최신 RNN은 새로운 정보가 들어올 때만 내부 상태를 업데이트하기 때문에, 더 오랜 시간 동안 사고하는 데에는 적합하지 않다. 더불어, 반복적인 깊이(recurrent depth)를 지원하는 RNN조차도 명시적인 검증을 위한 메커니즘이 부족하여, 시스템 2 사고를 위한 활용이 제한적이다.

AI의 주요 목표 중 하나는 어떤 종류의 문제에 대해서든 스스로 사고하는 법을 학습할 수 있는 시스템을 만드는 것이다. 이러한 목표는 궁극적으로 다음과 같은 핵심 질문으로 이어진다: **"시스템 2 사고를 개발하는 데 오직 비지도 학습만으로 충분할 수 있는가?"** 이러한 능력이 가능해진다면, 현재의 시스템 2 사고 접근법을 어떤 문제, 어떤 모달리티에도 일반화할 수 있으며, 외부의 인간, 보상, 또는 모델 기반 감독(supervision)에 의존하지 않아도 될 것이다.

본 연구에서는 위 핵심 질문에 대해 긍정 답변을 실증적으로 입증한다. 그러나 이러한 형태의 비지도 학습 기반 시스템 2 사고가 자연스럽게 나타나는 것을 방해하는 기존 모델 아키텍처의 여러 한계점이 존재함도 함께 확인했다. 특히 인간의 시스템 2 사고의 특성과 현재의 모델링 패러다임을 비교해보면(Figure 1, Table 1 참조), 다음과 같은 세 가지 핵심적인 차이점, 즉 시스템 2 사고의 세 가지 주요 측면(Facets)이 존재함을 알 수 있다:

측면 1: 계산 자원의 동적 할당- 인간은 과제의 난이도에 따라 서로 다른 정도의 노력을 자연스럽게 할당하며, 이는 심리학 및 신경과학 연구에서도 널리 알려져 있다. 인간이 직면하는 과제의 난이도는 매우 다양하기 때문에, 과제에 할당하는 계산 자원의 양을 조절하는 능력은 성공의 핵심 요소이다. 예) 직업을 바꿀지 말지를 결정하는 일은 일반적으로 점심 메뉴를 고르는 것보다 훨씬 더 많은 시간을 들여 고민한다.

측면 2: 연속 상태 공간에서의 불확실성 모델링- 더 오래 사고하는 것이 성능 향상에 중요하긴 하지만, 인간은 결정에 도달하기 전에 자신이 얼마나 불확실한지를 고려한다. 언어 영역에서는 LLM이 토큰 수준의 확률을 통해 이러한 불확실성을 시뮬레이션한다. 하지만 시각과 같은 연속적인 상태공간에서는, 벡터 양자화 같은 이산화 기법이나, ELBO 같은 pseudo losses/objectives 함수를 사용하지 않는 한, 트랜스포머, RNNs, 또는 확산 모델 같은 주요 모델들의 표준 구현은 일반적으로 신뢰도 높은 불확실성 추정을 제공하지 못한다. 이에 반해, EBM은 예측의 정확한 확률을 모델링하지 않더라도 상대적인 정규화되지 않은 확률을 통해 자연스럽게 불확실성을 모델링할 수 있다(Figure 3 참조). 현실 세계는 본질적으로 예측 불가능한 요소들로 가득 차 있으며, 예를 들어 주차된 차량 뒤에서 갑자기 보행자가 튀어나오는 상황처럼 불확실성이 중요한 경우가 많다. 따라서, 예측에서 불확실성을 표현할 수 있는 능력은 신중한 판단을 위해 필수적이며, 인간에게는 자연스럽게 지니고 있는 능력이다.

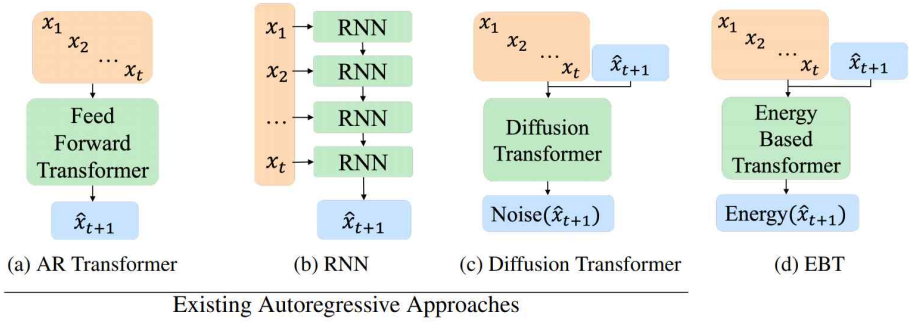


Figure 1: **Autoregressive Architecture Comparison.** (a) Autoregressive (AR) Transformer is the most common, with (b) RNNs becoming more popular recently [22, 23]. (c) Diffusion Transformers [26] (which are often bidirectional but can also be autoregressive) are the most similar to EBT, being able to dynamically allocate computation during inference, but predict the noise rather than the energy [27, 28]. Consequently, diffusion models cannot give unnormalized likelihood estimates at each step of the thinking process, and are not trained as explicit verifiers, unlike EBTs.

Table 1: **Architectures and Cognitive Facets.** For each prediction, Feed Forward (FF) Transformers and RNNs generally[2] have a finite amount of computation. While diffusion models have potentially more computation during inference by increasing the number of denoising steps, they do not learn to explicitly verify or estimate uncertainty for predictions. EBMs can use a dynamic amount of computation during inference by iterating for any number of steps, and give an energy scalar that can be used to evaluate uncertainty and verify the strength of predictions.

Architecture	Dynamic Compute Allocation (Facet 1)	Modeling Uncertainty (Facet 2)	Prediction Verification (Facet 3)
FF Transformers	✗	✗	✗
RNNs	✗	✗	✗
Diffusion Transformers	✓	✗	✗
EBTs	✓	✓	✓

측면 3: 예측의 검증- 계산 자원의 할당이나 불확실성 모델링 외에도, 효과적인 사고는 예측을 검증하는 능력으로부터도 큰 이점을 얻는다. 이 능력은 언제 사고를 멈출지 또는 가장 정확한 예측을 선택할지에 대한 결정을 내리는 데 중요한 역할을 하며, 실제로, 예측 검증이 인간 사고의 핵심 구성 요소이다. 사실, 해결책을 검증하는 것이 해결책을 생성하는 것보다 쉬운 문제이다. 이는 검증기가 생성기보다 더 잘 일반화할 수 있다는 의미이다. 명시적인 검증기를 훈련(학습)시키면, 사고 과정의 각 단계에서 현재 예측의 품질을 추정할 수 있다. 이러한 능력은 다음과 같은 동적 추론 시간(inference-time)의 행동을 가능하게 한다: 예측이 정확하다고 판단되면 조기에 중단, 문제가 어려울 경우 더 많은 계산 자원 할당. 이러한 행동은 직관적으로 과제의 난이도에 따라 자원을 조절하는 것과 유사하다. 또한, 명시적인 검증기 훈련을 통해, 몬테카를로 트리 탐색이나 다수 샘플링 후 최적 예측 선택과 같은 능력도 구현할 수 있다.

앞에서 설명한 모든 측면을 달성하기 위해, 사고를 학습된 검증기를 기준으로 한 최적화 과정으로 바라보고자 한다. 이 검증기는 입력과 후보-예측 간의 호환성을 평가한다. 좀 더 구체적으로 말하면, EBM을 훈련시켜, 가능한 모든 입력-예측 쌍에 대한 에너지 지형을 학습한다. 이때 낮은 에너지는 높은 호환성을 의미한다. 그런 다음, 사고 과정은 무작위 초기 예측에서 시작하여, 에너지를 점차 최소화하는 방식으로 예측을 반복적으로 정제해 나가는 것으로 정의된다(Figure 3). 학습된 에너지 지형을 최적화하는 이 방식은 자연스럽게 계산 자원의 동적 할당(Facet 1)을 가능하게 하며, 모델이 어려운 문제일수록 더 많은 반복 단계를 수행하여 더 오래 사고할 수 있다. 또한, 사고의 각 단계에서 EBM은 forward pass 상에서 검증기 역할을

하여, 현재 문맥과 예측 사이의 호환성을 나타내는 에너지를 출력한다. 이 에너지는 정규화되지 않은 확률로 작동함과 동시에, 예측이 올바른지를 판단하는 점수 역할도 하므로, 이는 불확실성 모델링(Facet 2)과 예측 검증(Facet 3)을 직접적으로 해결해준다(Figure 3).

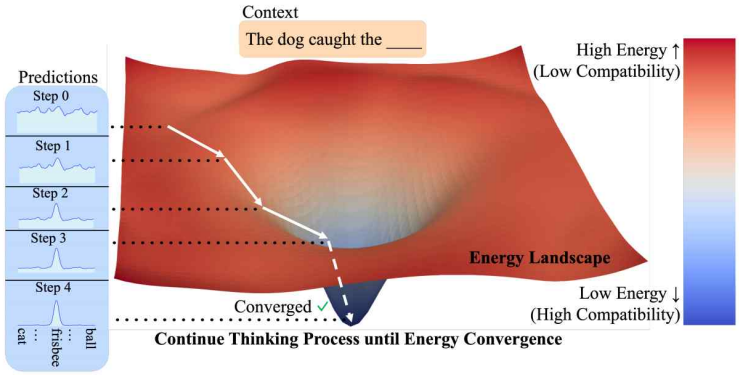


Figure 3: **Thinking Process Visualization.** A learned energy landscape and its optimization through gradient descent, interpreted as a thinking process. In this example, the model predicts a distribution over text tokens, progressively shifting from an initial random distribution toward the target distribution. At each step, the EBM assigns an energy scalar indicating how **compatible** the current prediction is with the context, visualized as the landscape's height (Facet 3). This scalar's convergence allows the model to determine whether the prediction is adequate or if further thinking is necessary. Uncertainty (Facet 2) can be represented by landscapes that are harder to optimize or by landscapes with many local minima, allowing the model to know when it requires more steps to think (Facet 1). Adapted from [57].

추론 시점 에너지 최소화 관점에서 사고를 바라보는 것은 유망한 접근이지만, EBM은 전통적으로 확장성에 어려움을 겪어왔으며, 공개적으로 알려진 기반(foundation) EBM은 아직 존재하지 않는다. 이러한 한계는 EBM의 훈련 안정성 문제와 긴 학습 시간 때문인 것으로 알려져 있다. 이러한 문제를 해결하고 확장 가능한 EBM 학습 패러다임을 확립하기 위해, EBT를 제안한다. 이는 EBM을 위한 트랜스포머 기반 구현 형태이다.

시스템 2 사고가 일반화에 미치는 효과를 분석한 결과, 두 가지 현상이 일관되게 관찰되었다. 첫째, 동일하거나 낮은 사전 학습 성능을 가진 경우에도, EBT는 시스템 2 사고를 수행함으로써 기존 모델보다 추론 단계에서 더 우수한 성능을 발휘하여, 시스템 2 사고의 중요성과 EBT의 일반화 능력을 입증하였다. 둘째, EBT의 시스템 2 사고는 분포에서 더 멀리 떨어진(out-of-distribution) 데이터일수록 더 큰 성능 향상을 보여주는데, 이는 인간이 어려운 미지의 과제에 대해 시스템 2 사고를 사용하는 심리학적 관찰과도 일치한다.

Language Models: TextGrad와 Agentic AI의 미래

취지: LLM(Large Language Model)은 현재 AI 분야에서 가장 큰 관심을 끌고 있다. 그리고, LLM을 활용한 에이전트 AI의 개발에도 많은 노력을 기울이고 있다. 이러한 추세에 참고할 연구 결과 두가지를 조사하였다. 첫 번째는 LLM이 생성한 자연어 피드백을 역전파하여 AI 시스템을 개선하는 TextGrad라는 범용 프레임워크이다. 이는 다양한 작업에서 생성형 AI의 자동 최적화를 가능하게 해준다. 그리고, 두 번째 연구는 SLM(Small Language Models)이 소수의 특화된 작업을 변동 없이 반복적으로 수행하는 에이전트 AI 개발의 미래 방향임을 주장한다. 즉, SLM이 에이전틱 시스템의 많은 활용에서 충분한 성능을 지녔으며, 본질적으로 더 적합하고, 필연적으로 더 경제적이기 때문에 에이전틱 AI의 미래라고 주장한다.

조사자료:

M. Yuksekgonul et al., “Optimizing generative AI by backprpagating lenguage model feedback,” Nature, Vol. 639, pp. 609-616, March 2025.

M. Yuksekgonul et al., “TextGrad: Automatoc ‘Differentiation’ via Text,” https://doi.org/10.48550/arXiv.2406.07496

P. Belcak et al., “Small Language Models are the Furture of Agentic AI,” https://doi.org/10.48550/arXiv.2506.02153

TextGrad: “Optimizing generative AI by backpropagating language model feedback”

M. Yuksekgonul et al., Nature, Vol. 639, pp. 609-616, March 2025.

요약: 최근 AI의 혁신은 검색 엔진이나 시뮬레이터와 같은 특화 도구들뿐 아니라, 여러 대규모 언어 모델(LLM)을 조율하는 시스템에 의해 점점 더 주도되고 있다. 지금까지 이러한 시스템은 주로 도메인 전문가가 직접 설계하고 휴리스틱을 통해 조정되어 왔으며, 자동으로 최적화되지 못한다는 점이 발전 가속화의 큰 과제로 남아 있었다. 인공 신경망의 발전 또한 과거에 비슷한 과제에 직면했으나, 역전파와 자동 미분의 등장으로 최적화가 자동화되면서 해당 분야는 크게 변모했다. 이와 유사하게, LLM이 생성한 피드백을 역전파하여 AI 시스템을 개선하는 TextGrad라는 범용 프레임워크를 소개한다. TextGrad는 자연어 피드백을 활용하여 시스템의 어떤 부분이든 비판(critique)과 개선 제안을 수행함으로써, 다양한 작업에서 생성형 AI 시스템의 자동 최적화를 가능하게 한다. TextGrad는 과학자와 엔지니어가 영향력이 막대한 생성형 AI 시스템을 손쉽게 개발할 수 있도록 한다.

1 서론

대규모 언어 모델(LLM)은 혁신적인 인공지능(AI) 시스템이 구축되는 방식을 변화시키고 있다. 최신 AI 응용 프로그램은 점점 더 복합 시스템 형태로 작동하며, LLM 기반 에이전트에서 시뮬레이터나 웹 검색 엔진과 같은 특화 도구에 이르기까지 여러 정교한 구성 요소들을 조율한다. 예를 들어, LLM들이 기호적 해석기(symbolic solver)와 상호작용하는 시스템은 올림피아드 수준의 수학 문제를 풀 수 있으며, LLM이 검색 엔진과 코드 해석 도구를 활용하는 시스템은 인간 경쟁 프로그래머에 필적하는 성과를 내고 실제 GitHub 이슈를 해결하기도 한다. 이러한 진보에도 불구하고, 많은 돌파구들은 여전히 도메인 전문가가 직접 설계한 시스템에서 나온 결과이다. 따라서 LLM을 활용한 복합 시스템을 구축하고 향후 돌파구를 가속화하기 위해서는 자동 최적화 알고리즘을 개발하는 것이 가장 중요한 과제 중 하나이다.

지난 15년 동안 AI의 많은 발전은 신경망과 미분 가능한 최적화(differentiable optimization)에 의존해 왔다. 신경망의 구성 요소들은 행렬 곱셈과 같은 미분 가능한 함수들을 통해 상호 작용한다. 따라서 각 매개변수를 조정하여 모델을 개선할 수 있는 방향을 제공하는 수치적 기울기를 역전파 방식으로 각 구성 요소에 전달하는 것이 AI 모델을 학습하는 자연스러운 방법이었다. 그러나 새롭게 등장한 생성형 AI 시스템에서는 미분 가능한 최적화가 쉽게 작동하지 않는다. 이러한 시스템은 종종 자연어 상호작용, 블랙박스 형태의 LLM, 또는 외부 도구들을 포함하기 때문에 수치적 기울기를 역전파하기가 어렵기 때문이다.

이 과제를 해결하기 위해, 텍스트를 통한 자동 ‘미분’을 수행하는 TextGrad를 소개한다. 신경망이 수치 값을 통해 소통하는 반면, 현대 AI 시스템은 텍스트, 코드, 이미지와 같은 비정형 데이터로 상호 작용한다. TextGrad에서는 각 AI 시스템을 계산 그래프로 변환하여, 그 구성 요소들이 이러한 풍부한 비정형 변수를 복잡한(반드시 미분 가능하지는 않은) 함수들을 통해 주고받도록 한다. 여기서 ‘미분’을 은유적으로 사용하며, LLM이 각 변수를 어떻게 수정해야 전체 시스템이 개선될 수 있는지를 설명하는 유익하고 해석 가능한 자연어 비평 형태의 피드백(‘텍스트 기울기’라 부름)을 제공한다(그림 1a,b,c). 텍스트 기울기는 LLM API 호출, 시뮬레이터, 외부 수치 해석기와 같은 임의의 함수를 거쳐 전파될 수 있다. 이러한 텍스트 피드백은

기반 신경망의 매개변수를 직접 겨냥하지 않으므로 블랙박스 API와도 호환된다.

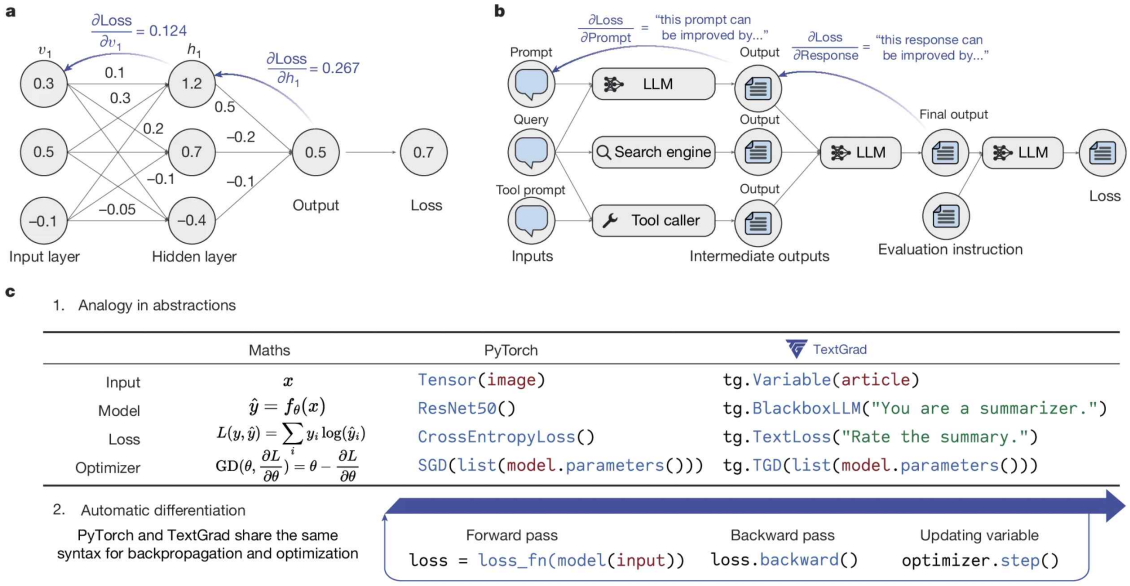


그림 9

TextGrad의 강력함을 질문-응답 벤치마크에서부터 방사선 치료 계획 최적화와 분자 생성에 이르기까지 다양한 분야에서 입증하였다. 이 모든 작업은 서로 다른 과제에 맞게 수정할 필요 없이 단일 프레임워크로 수행된다. LLM은 분자 수정 제안, 다른 LLM을 위한 프롬프트, 코드 snippet과 같이 다양한 영역의 변수들에 대해 풍부하고 가독성 있으며 표현력 있는 자연어 피드백을 제공할 수 있다. 이 프레임워크는 최신 LLM이 최적화 대상 시스템의 개별 구성 요소와 하위 과제에 대해 추론할 수 있다는 전제를 바탕으로 구축되었다. 이 결과는 TextGrad가 생성형 AI 시스템과 그 출력물을 자동으로 개선할 수 있는 가능성을 보여준다.

2 TextGrad를 활용한 LLM 피드백의 역전파

먼저 두 번의 LLM 호출로 이루어진 시스템 예시를 통해 TextGrad가 어떻게 동작하는지를 설명한다. 다음 시스템을 고려하자:

$$\text{Prediction} = \text{LLM}(\text{Prompt}+\text{Question}) \quad (1)$$

$$\text{Evaluation} = \text{LLM}(\text{Evaluation Instruction} + \text{Prediction}) \quad (2)$$

여기서 +는 두 문자열의 연결(concatenation)을 의미하고, LLM(x)는 입력 x를 언어 모델에 프롬프트로 제공하여 그 응답을 받는 과정을 뜻한다. 이러한 구조는 언어 모델이 평가 목적으로 사용될 때, 예를 들어 LLM-as-a-judge와 같은 경우에 자주 활용된다.

전통적인 자동 미분에서는 연쇄 법칙을 사용해 기울기를 계산함으로써, 시스템의 수치적 변수를 어떻게 수정해야 목표 함수에 비추어 시스템을 개선할 수 있는지를 결정한다. 이에 비해 TextGrad는 시스템 내 비정형 변수를 수정할 수 있도록 안내하는 텍스트 형태의 피드백을 생성한다. TextGrad는 수치적 기울기를 계산하지 않고, 대신 목표 함수에 따라 시스템을 개선할 수 있는 구체적인 수정 방향을 제안하는 비평을 제공한다.

AI 시스템을 개선(학습)하기 위해, 우리는 역전파 알고리즘의 유사 형태를 구현한다(그림 2a). 여기서, 순방향 함수가 LLM 호출일 때 텍스트 기울기 연산자를 ∇LLM이라 한다. ∇LLM을 구현하는 유연한 방법은 그림 2b에 제시되어 있다. 특히 이 함수는 "예측은 …를 통

해 개선될 수 있습니다"와 같은 비평을 반환하며, 이는 수치적 기울기와 유사하게 목표 함수에 비추어 전체 시스템을 개선하기 위해 변수를 어떻게 수정해야 하는지를 설명한다. 이 맥락에서 수정 가능한 구성 요소는 프롬프트, 중간 출력, 혹은 최종 예측이 될 수 있다. 그림 2a에서 프롬프트로 피드백을 역전파하기 위해, 우리는 먼저 평가를 통해 예측 변수에 대한 피드백을 수집한다. 그런 다음, 이 피드백과 LLM(Prompt + Question) 호출을 바탕으로 프롬프트에 대한 피드백을 수집한다. 보다 일반적으로는, 한 변수의 모든 후속 요소들로부터 얻은 피드백을 활용해 이 절차를 적용한다.

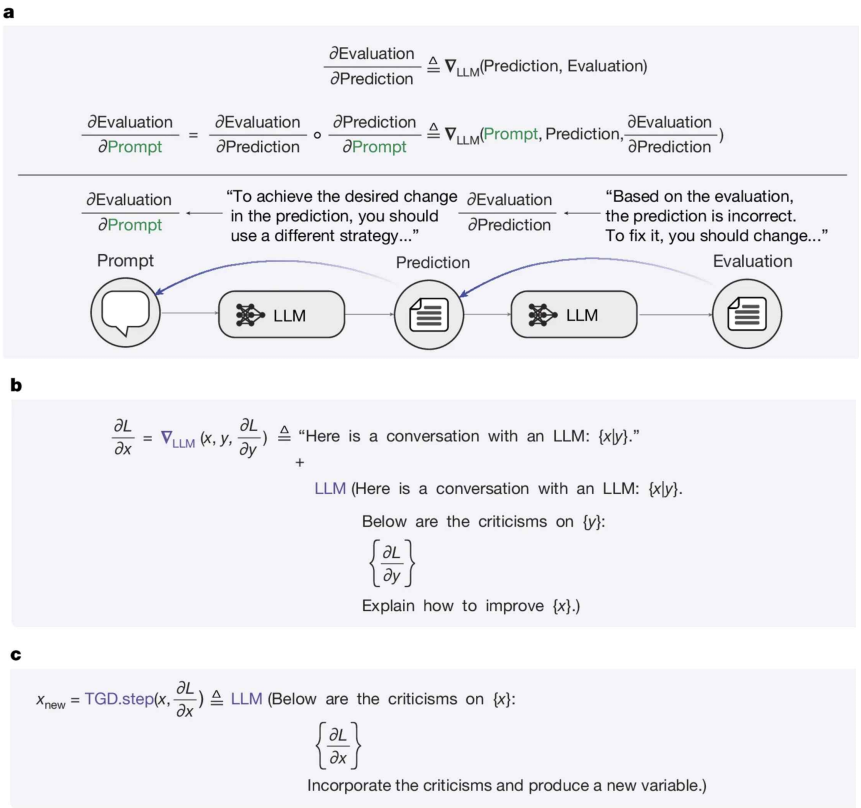


그림 2

수치적 경사 하강법에서는 매개변수 θ를 손실 함수에 대한 기울기와 결합하여, 뺄셈을 통해 $\theta_{\text{new}}(\theta, \frac{\partial L}{\partial \theta}) = \theta - \frac{\partial L}{\partial \theta}$ 와 같이 갱신한다. 이 경사 기반 최적화의 유추를 이어서, 텍스트 경사 하강법(Textual Gradient Descent, TGD)을 사용한다. 여기서는 변수를 피드백을 이용해 갱신한다. 그림 2c에서는 TGD를 구현하는 구체적인 방법을 보여준다. 이 단계에서는 텍스트 기울기를 활용해 목표 함수에 따라 시스템을 개선할 수 있는 방식으로 구성 요소들을 수정한다.

자동 미분에서는 목표 함수가 일반적으로 평균 제곱 오차와 같은 미분 가능한 함수이다. 반면, TextGrad에서는 목표 함수가 복잡하고 미분 불가능하거나 블랙박스 형태일 수 있으며, 함수의 정의역과 공역이 비정형 데이터일 수도 있다. 이러한 선택은 프레임워크에 중요한 범용성과 유연성을 부여한다. 예를 들어, 목표 함수는 언어 모델을 프롬프트로 호출하여 얻거나, 단위 테스트를 실행하는 코드 해석기 출력, 혹은 분자 시뮬레이션 엔진의 출력으로 얻을 수 있다. LLM을 사용한 단순한 목표 함수는 다음과 같이 구성될 수 있다:

Loss(question,solution))=LLM(""Below is a multi-choice question and a solution:
{question,solution} Your job is to investigate the solution,

Critically go through reasoning steps, and
see if the prediction could be incorrect.")

여기서 이 평가 신호를 활용하여 코드 스니펫을 최적화할 수 있으며, 이는 LLM이 스스로 비평하고 개선할 수 있는 잠재력에 기반한다.

TextGrad를 두 가지 넓은 범주의 과제를 해결하는 데 사용한다. 첫째, 테스트 시점 최적화(test-time optimization)에서는 문제의 해답—예를 들어 코드 스니펫, 질문의 정답, 또는 분자—을 최적화 변수로 직접 다룬다. 예를 들어 앞서 언급한 질문-응답 목표 함수에서는 테스트 시 개선하고자 하는 해답이 존재한다. 우리의 프레임워크는 이 해답 변수에 대한 텍스트 기울기를 생성하고, 이를 반복적으로 적용하여 해답 변수를 개선한다. 둘째, 프롬프트 최적화에서는 목표가 특정 과제에 대해 LLM의 여러 쿼리 성능을 향상시키는 프롬프트를 찾는 것이다. 예를 들어 수학적 추론 문제에서 LLM의 성능을 향상시키는 시스템 프롬프트를 찾고자 할 수 있다. 특히 프롬프트 최적화에서는 테스트 시점 최적화와 달리, 특정 문제에 국한되지 않고 새로운 문제에서도 잘 일반화되도록 시스템 프롬프트를 찾는 것이 목표이다. 이 두 가지 유형의 문제는 모두 프레임워크를 수작업으로 설계하지 않고 해결할 수 있다.

TextGrad의 유연성을 다섯 가지의 다양한 응용 분야에서 입증하였다. (1) 온라인 프로그래밍 학습 플랫폼인 LeetCode의 난이도 높은 문제를 해결하기 위해 코드 스니펫을 최적화한다. (2) 과학적 질문의 해답을 최적화한다. (3) LLM의 추론을 안내하기 위한 프롬프트를 최적화한다. (4) 전립선암 환자의 방사선 치료 계획을 최적화한다. (5) 여러 변수와 긴 체인을 포함하는 복합 시스템을 최적화한다. 또한, 보충 자료(Supplementary Note 8.1)에서는 분자 특성 최대화를 위한 화학 구조 최적화 사례를 논의한다. 58개의 목표 단백질 데이터셋에서, TextGrad는 기존의 분자 최적화 방법과 비교하여 리간드 설계 성능이 향상됨을 보여준다 (Extended Data 그림 1 및 2).

4 논의(Discussion)

TextGrad는 세 가지 원칙에 기반한다. (1) 특정 응용 분야에 맞춰 수작업으로 설계된 것이 아니라, 범용적이고 성능이 우수한 프레임워크이다. (2) 사용이 용이하며(PyTorch 추상화를 반영), 지식 이전을 지원한다. (3) 완전히 오픈소스이다. TextGrad를 통해 코드 최적화 및 박사 수준 질문 응답에서 SOTA 성과를 달성했으며, 프롬프트 최적화를 수행하고, 분자 설계와 치료 계획 최적화에서 개념 검증 결과를 보였다. 이전 연구들은 LLM을 프롬프트 엔지니어링과 자기 개선(self-improvement)을 위한 최적화 도구로 활용하는 방법을 탐구했다. 이러한 프레임워크는 프롬프트 최적화와 같은 특정 영역에서 LLM의 능력을 입증했지만, 본 연구에서는 자연어 피드백을 역전파하는 독창적인 방법에 기반한 범용 프레임워크를 제안한다.

TextGrad가 과학적 발견에서 iterative 과정을 가속화하고 공학적 연구의 생산성을 높이는 데 사용될 수 있기를 기대한다. 그리고, 텍스트 피드백을 역전파하는 알고리즘에 대한 설계 공간(design space)이 매우 넓다. 수치 최적화, 자동 미분과 TextGrad 사이에 많은 연관성을 도출할 수 있다 생각한다. AI 시스템이 개별 모델 학습에서 여러 상호작용하는 LLM 구성 요소와 도구를 포함한 복합 시스템 최적화로 전환됨에 따라, 새로운 세대의 최적화 도구가 필요하다. TextGrad는 LLM의 추론 능력과 역전파의 분해 가능 효율성을 결합하여 AI 시스템을 최적화할 수 있는 범용 프레임워크를 제공한다.

Small Language Models are the Future of Agentic AI

P. Belcak et al., https://doi.org/10.48550/arXiv.2506.02153

요약: 대규모 언어 모델(LLMs)은 광범위한 작업에서 인간에 가까운 성능을 보이고, 일반적인 대화를 나눌 수 있는 성능으로 높이 평가받고 있다. 한편으로, 에이전틱(agentic) AI 시스템의 부상은, 언어 모델이 소수의 특화된 작업을 변동 없이 반복적으로 수행하는 수많은 응용 사례를 보여주고 있다. 본 논문에서는 소형 언어 모델(SLMs)이 이러한 에이전틱 시스템의 많은 활용에서 충분한 성능을 지녔으며, 본질적으로 더 적합하고, 필연적으로 더 경제적이기 때문에 에이전틱 AI의 미래라고 주장한다. 이는 SLM이 현재 보여주고 있는 역량 수준, 에이전틱 시스템의 일반적 아키텍처, 그리고 언어 모델 배포의 경제성을 근거로 제시한다. 또한, 범용적인 대화 능력이 본질적으로 필요한 상황에서는 이기종 에이전틱 시스템(heterogeneous agentic systems, 즉 여러 다른 모델을 호출하는 에이전트)이 자연스러운 선택이다. 여기서는 SLM을 에이전틱 시스템에 도입하는 데 존재하는 잠재적인 장벽을 논의하고, LLM을 SLM 기반 에이전트로 전환하는 일반적 알고리즘을 제시한다. 이 논문의 입장은 가치 선언(value statement)으로서, LLM에서 SLM으로의 부분적 전환만으로도 AI 에이전트 산업에 미칠 운영적·경제적 영향의 중요성을 강조한다. 이를 통해 AI 자원의 효율적 활용에 대한 논의를 촉발하고, 오늘날 AI 비용 절감 노력을 진전시키고자 한다.

1 서론

조사에 따르면 대형 IT 기업의 절반 이상이 AI 에이전트를 적극적으로 활용하고 있으며, 그 중 21%는 지난 1년 안에 도입한 것이다. 사용자 측면 뿐만 아니라 시장에서도 AI 에이전트의 경제적 가치가 크게 주목받고 있다. 2024년 말 기준, 에이전틱 AI 분야는 20억 달러 이상의 스타트업 투자를 유치했으며, 52억 달러 규모로 평가되었고, 2034년까지 거의 2천억 달러에 이를 것으로 전망된다. 이처럼 AI 에이전트가 현대 경제에서 중요한 역할을 담당할 것이다.

대부분의 현대 AI 에이전트를 구동하는 핵심 구성 요소는 LLM이다. 에이전트가 어떤 도구를 언제, 어떻게 사용할지에 대해 전략적으로 결정하고, 과제를 완수하기 위해 필요한 작업의 흐름을 제어하며, 필요하다면 복잡한 과제를 관리 가능한 하위 과제로 나누고 행동 계획 및 문제 해결을 위한 추론을 수행할 수 있게 해주는 기반 지능을 제공하는 것이 바로 LLM이다. 일반적인 AI 에이전트는 선택한 LLM API 엔드포인트와 단순히 통신하면서, 이러한 모델을 호스팅하는 중앙화된 클라우드 인프라에 요청을 보내는 방식으로 동작한다.

LLM API 엔드포인트는 단일 범용 LLM을 활용해 방대한 양의 다양한 요청을 처리하도록 특별히 설계되어 있다. 이러한 운영 모델은 업계에 깊이 뿌리내려 있으며 —그 정도가 상당하여 막대한 자본 투자의 기반이 되고 있다. 예를 들어, 2024년 에이전틱 애플리케이션을 뒷받침하는 LLM API 서비스 시장 규모는 56억 달러로 추정되었으나, 같은 해 이를 호스팅하는 클라우드 인프라에 대한 투자는 570억 달러로 급증했다. 투자 규모와 시장 규모 간의 10배 격차는 업계에서 당연하게 받아들여지고 있는데, 이는 이러한 운영 모델이 본질적으로 크게 변하지 않은 채 업계의 초석으로 남을 것이며, 이는 대규모 초기 투자가 3~4년 내 전통적인 소프트웨어 및 인터넷 솔루션과 견줄 만한 수익을 가져올 것이라 기대하기 때문이다.

여기서는 표준 운영 모델의 지배적 위치를 인정하면서도, 그중 한 가지 측면 —즉, 상대적으로 단순한 요청임에도 불구하고 에이전트가 언어 지능에 접근하기 위해 일반 범용 LLM에 의존하는 관행— 을 문제 삼는다. 소규모 LM이야말로 에이전틱 AI의 미래라는 입장을 제시(2

장), 논증(3장), 옹호(4장)한다. 다만, 현재 상황이 그와 반대로 전개되는 원인으로서 기업의 투자 및 이미 굳어진 관행(5장)을 인정한다. 이에 대한 대안으로, 에이전틱 애플리케이션을 LLM에서 SLM으로 이전하기 위한 전환 알고리즘을 제시한다(6장).

2 입장(Position)

2.1 정의(Definitions)

이 논문의 입장을 구체화하기 위해, 다음과 같은 작업 정의(working definitions)를 사용한다:

WD1: SLM은 일반적인 소비자용 전자기기에 탑재될 수 있으며, 단일 사용자의 에이전틱 요청을 처리할 때 실용적으로 충분히 낮은 추론 지연으로 동작하는 언어 모델을 의미한다.

WD2: LLM은 SLM이 아닌 언어 모델을 의미한다.

2025년 기준으로, 대체로 100억 개 미만의 파라미터를 가진 모델을 SLM으로 간주하는 데 무리가 없다고 본다. 에이전트와 에이전틱 시스템이라는 용어를 서로 바꿔 사용하되, 전체적으로 일정한 자율성을 가진 소프트웨어 자체를 강조할 때는 전자를(예: “인기 있는 코딩 에이전트에서 볼 수 있듯이”), 에이전틱 애플리케이션을 구성 요소들의 합으로서 시스템적 측면을 강조할 때는 후자를 선호한다(예: “에이전틱 시스템의 모든 언어 모델이 SLM으로 대체 가능한 것은 아니다”). 간결함을 위해, 에이전틱 애플리케이션의 기반으로서 언어 모델에 집중하며, 비전-언어 모델은 명시적으로 다루지 않는다. 그러나 이 논문의 입장과 대부분의 논거가 비전-언어 모델에도 쉽게 확장될 수 있음을 덧붙여 언급한다.

2.2 주장(Statement)

SLM이 다음과 같다고 주장한다:

V1: 에이전틱 애플리케이션의 언어 모델링 과제를 처리하기에 본질적으로 충분한 성능을 갖추고 있으며,

V2: LLM보다 에이전틱 시스템에서 운영적으로 더 적합하고,

V3: 규모가 작다는 특성 덕분에, 에이전틱 시스템에서 대부분의 언어 모델 활용에 있어 범용 LLM 대비 필연적으로 훨씬 더 경제적이다.

그리고 이러한 V1-V3의 근거를 바탕으로, SLM이 에이전틱 AI의 미래라고 주장한다.

이 논문의 입장 표현은 의도적이다. 이 주장에서는, 자연스러운 우선순위를 따를 경우 SLM과 LLM 간의 차이로 인해 앞으로의 발전이 궁극적으로 필연적인 결과임을 전달하고자 한다.

2.3 상세 설명(Elaboration)

AI 에이전트 설계에서 LLM의 지배가 과도하며, 대부분의 에이전틱 사용 사례의 기능적 요구와 맞지 않다고 이 논문은 주장한다. LLM이 뛰어난 범용성과 대화 유창성을 제공하긴 하지만, 실제 배포된 에이전틱 시스템에서 수행되는 대부분의 하위 과제는 반복적이고 범위가 제한적이며 대화 중심이 아니다. 이러한 과제에는 효율적이고, 예측 가능하며, 비용이 낮은 모델이 필요하다. 이러한 맥락에서 SLM은 단순히 충분할 뿐만 아니라 더 적합하다. SLM은 낮은 지연(latency), 감소된 메모리와 연산 요구량, 훨씬 낮은 운영 비용을 제공하면서도, 제한된 영역 내에서 충분한 과제 수행 성능을 유지할 수 있다는 장점을 지녔다.

이 논문의 입장은 에이전틱 아키텍처 내 언어 모델 사용 패턴에 대한 실용적 관점에서 비롯된다. 이러한 시스템은 일반적으로 복잡한 목표를 모듈화된 하위 과제로 분해하며, 각 하위 과제는 특화되거나 파인튜닝된 SLM으로 안정적으로 처리된다. 모든 과제에 LLM을 고집하는 것은 계산 자원의 잘못된 배분을 반영하는 것으로, 규모가 커질수록 경제적으로 비효율적이고

환경적으로 지속 불가능하다고 주장한다. 또한, 일반적인 추론이나 개방형 대화가 필수적인 경우에는 이종(agentic) 시스템을 권장한다. 이 시스템에서는 기본적으로 SLM을 사용하고, LLM은 선택적이고 제한적으로 호출된다. 이러한 모듈식 구성—SLM의 정밀성과 효율성에 LLM의 범용성을 결합—은 비용 효율적이면서도 유능한 에이전트를 구축할 수 있게 한다.

궁극적으로, LLM 중심 구조에서 SLM 우선 구조로의 전환이 많은 사람들에게 단순한 기술적 개선을 넘어 휴메적(Humean) 도덕적 당위로 여겨진다는 점을 관찰한다. AI 커뮤니티가 증가하는 인프라 비용과 환경 문제에 직면함에 따라, 에이전틱 워크플로우에서 SLM의 사용을 채택하고 이를 표준화하는 것은 책임감 있고 지속 가능한 AI 배포를 촉진하는 데 중요하다.

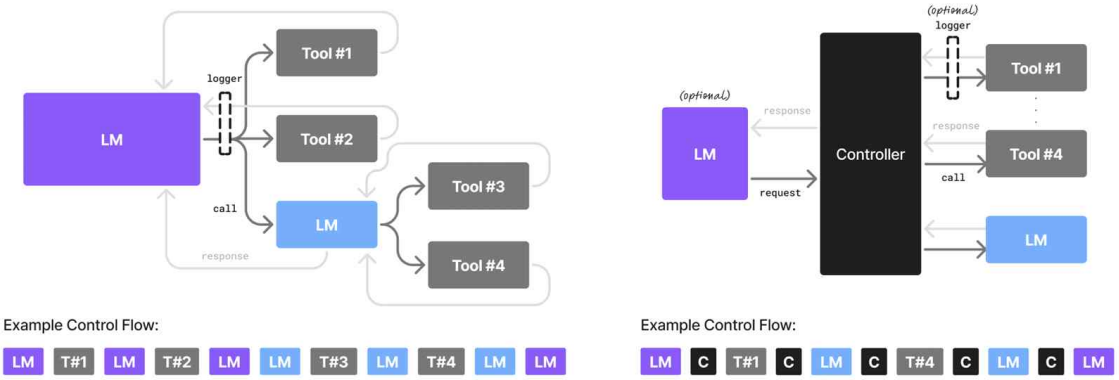


Figure 1: An illustration of agentic systems with different modes of agency. *Left: Language model agency.* The language model acts both as the HCI and the orchestrator of tool calls to carry out a task. *Right: Code agency.* The language model fills the role of the HCI (optionally) while a dedicated controller code orchestrates all interactions.

3 입장 논거(Position Arguments)

V1-V3 관점을 다음과 같은 배타적이지 않은 논거를 통해 지지한다.

3.1 SLM은 이미 에이전트에서 사용할 만큼 충분한 성능을 지녔다.

A1: SLM은 에이전틱 시스템에서 LLM을 대체할 만큼 충분히 강력하다(충분한 성능을 지녔다). 이 논거는 V1 관점을 지지한다.

지난 몇 년 동안, 소규모 언어 모델의 성능은 크게 향상되었다. 비록 LM 스케일링 법칙이 여전히 관찰되고 있지만, 모델 크기와 성능 사이의 스케일링 곡선은 점점 가팔라지고 있어, 최신 소규모 언어 모델의 성능이 이전 대규모 모델의 성능과 훨씬 가까워지고 있음을 의미한다. 실제로 최근 연구에서는, 잘 설계된 소규모 언어 모델이 이전에는 훨씬 큰 모델에서만 달성 가능했던 과제 수행 성능을 충족하거나 초과할 수 있음이 보여지고 있다.

아래 인용된 개별 연구들에서 대규모 모델과의 광범위한 비교가 수행되었지만, 벤치마크에서 평가된 모든 능력이 에이전틱 환경에서의 배포에 필수적인 것은 아니다. 여기서는 상식적 추론 능력(기본 이해의 지표), 도구 호출 및 코드 생성 능력(모델→도구/코드 인터페이스를 정확히 소통할 수 있는 능력의 지표; 그림 1 참조), 그리고 지시 수행 능력(코드←모델 인터페이스를 통해 올바르게 응답할 수 있는 능력)에 대한 적합성을 강조한다.

· **Microsoft Phi 시리즈:** Phi-2(27억 파라미터)는 상식적 추론 점수와 코드 생성 점수가 300억 모델과 비슷하면서도 15배 빠르게 동작한다. Phi-3 small(70억 파라미터)는 동일 세대의 700억 모델과 비슷한 수준의 언어 이해 및 상식적 추론 능력과 코드 생성 점수를 달성한다.

· **NVIDIA Nemotron-H 계열:** 2/4.8/9억 파라미터 하이브리드 Mamba-Transformer 모델

은 동일 세대의 300억 파라미터 LLM과 비슷한 수준의 지시 수행 및 코드 생성 정확도를 달성하면서도, 추론 FLOPs는 1/10 수준에 불과하다.

· **Huggingface SmolLM2 시리즈:** 1억 2,500만에서 17억 파라미터 규모의 컴팩트 언어 모델 SmolLM2 계열은 언어 이해, 도구 호출, 지시 수행 성능에서 동시대 140억 모델과 비슷한 수준을 보이며, 2년 전 700억 모델과도 맞먹는 성능을 보여준다.

· **NVIDIA Hymba-1.5B:** 이 Mamba-attention 하이브리드 SLM은 지시 수행 정확도에서 최고를 기록하며, 유사 크기 트랜스포머 모델 대비 토큰 처리량이 3.5배 높다. 지시 수행 능력에서는 130억 모델보다도 우수하다.

· **DeepSeek-R1-Distill 시리즈:** 15억~80억 파라미터 규모의 추론 모델 계열로, DeepSeek-R1이 생성한 샘플로 학습되었다. 상식적 추론 능력이 뛰어나며, 특히 DeepSeek-R1-Distill-Qwen-7B 모델은 Claude-3.5-Sonnet-1022나 GPT-4o-0513 같은 대형 독점 모델을 능가한다.

· **DeepMind RETRO-7.5B:** Retrieval-Enhanced Transformer(RETRO)는 75억 파라미터 모델로 방대한 외부 텍스트 데이터베이스가 보강되어 있으며, 언어 모델링 성능에서 GPT-3(1,750억)과 비슷한 수준을 달성하면서 파라미터 수는 25배 적다.

· **Salesforce xLAM-2-8B:** 80억 모델로 상대적으로 작은 규모임에도 불구하고 도구 호출에서 최첨단 성능을 달성하며, GPT-4o 및 Claude 3.5와 같은 최전선 모델을 능가한다.

특히, 경쟁력 있는 기본 성능 외에도 SLM의 추론 능력은 추론 시점에서 self-consistency, 검증자 피드백, 도구 보강 등을 통해 향상될 수 있다는 것이다. 예를 들어, Toolformer(67억)는 API 사용을 통해 GPT-3(1,750억)를 능가하며, 10~30억 규모 모델은 구조화된 추론(structured reasoning)을 통해 300억 이상 LLM과 수학 문제에서 맞먹는 성능을 보였다.

요약하면, 최신 학습, 프롬프트 기법, 에이전틱 보강 기술을 활용하면 제약 요인은 파라미터 수가 아니라 성능이다. SLM은 현재 상당수의 에이전틱 호출에서 충분한 추론 능력을 제공하며, 단순히 실행 가능한 수준을 넘어 모듈식이고 확장 가능한 에이전틱 시스템에서는 LLM보다 상대적으로 더 적합하다.

3.2 SLM은 에이전틱 시스템에서 더 경제적이다

A2: SLM은 에이전틱 시스템에서 LLM보다 더 경제적이다. 이 논거는 V3 관점을 지지한다.

소규모 모델은 비용 효율성, 적응성, 배포 유연성 측면에서 상당한 이점을 제공한다. 이러한 장점은 전문화와 반복적 개선이 중요한 에이전틱 워크플로우에서 특히 가치가 있다. 3.1절에서는 나열된 SLM과 관련 LLM 간의 여러 효율성 비교를 상세히 다루었다. 여기서는 논거 A2를 뒷받침하기 위해 보다 포괄적인 그림을 제시한다.

·**추론 효율:** 70억 파라미터 SLM을 운영하는 비용은 700~1,750억 파라미터 LLM 대비 10~30배 저렴하다(지연 시간, 에너지 소비, FLOPs 기준), 덕분에 대규모 환경에서도 실시간 에이전틱 응답이 가능하다. NVIDIA Dynamo와 같은 최신 추론 운영체제는 클라우드 및 엣지 배포에서 고속, 저지연 SLM 추론을 명시적으로 지원한다. 또한 SLM은 GPU나 노드 간 병렬화가 측소되기에, 서빙 인프라의 유지보수 및 운영 비용도 낮다.

·**파인튜닝 유연성(Fine-tuning agility):** SLM의 파라미터 효율적 파인튜닝(예: LoRA, DoRA) 또는 전체 파라미터 파인튜닝은 몇 시간의 GPU 작업만으로 가능하여, 기능을 추가, 수정, 또는 특화 시키는 작업을 몇 주가 아닌 하룻밤 만에 수행할 수 있다.

·**엣지 배포:** ChatRTX와 같은 온디바이스 추론 시스템의 발전은 소비자용 GPU에서 SLM을

로컬 실행할 수 있음을 보여주며, 이를 통해 실시간 오프라인 에이전틱 추론이 가능하고, 지연 시간은 줄이며 데이터 통제는 강화할 수 있다.

·**파라미터 활용:** 처음 보면, LLM은 방대한 양의 매개변수를 포함한 거대한 단일체처럼 작동하여, 압축된 정보 덩어리를 출력으로 생성하는 것처럼 보인다. 그러나 더 면밀히 들여다보면, 이러한 시스템을 통과하는 신호의 상당 부분은 희소하며, 어떤 단일 입력에 대해서도 매개변수의 일부분만이 작동한다는 것을 알 수 있다. 이러한 특성이 SLM에서는 더 완화되어 나타난다는 점은, SLM이 출력에 실질적 영향을 주지 않으면서도 추론 비용에 기여하는 매개변수의 비율이 더 작다는 점에서 더 효율적일 수 있음을 시사한다.

모듈식 시스템 설계: 크기가 다양한 여러 모델을 활용하는 접근 방식이 실제 에이전틱 과제의 이질성과 잘 맞으며, 이미 주요 소프트웨어 개발 프레임워크에 서서히 도입되고 있다. 또한, 에이전트 맥락에서 새롭게 발견된 이러한 모듈화 감각은 새로운 기술을 쉽게 추가하고 변화하는 요구사항에 적응할 수 있게 하며, 언어 모델 설계에서의 모듈화 추구와도 일치한다.

앞서 언급한 "레고(Lego)식" 에이전틱 지능 구성-거대 단일 모델을 확장하는 대신, 작고 전문화된 모델을 추가하여 수평적으로 확장하는 방식-은 더 저렴하고, 디버깅이 빠르며, 배포가 용이하고, 실제 에이전트의 운영 다양성과도 더 잘 맞는 시스템을 만든다. 여기에 도구 호출, 캐싱, 세분화된 라우팅을 결합하면, SLM 우선 아키텍처가 비용 효율적이고, 모듈식이며, 지속 가능한 에이전틱 AI로 나아가는 최적의 경로를 제공하는 것으로 보인다.

3.3 SLM은 더 유연하다

A3: SLM은 LLM에 비해 운영적 유연성이 더 크다. 이 논거는 V2와 V3 관점을 지지한다.

작은 규모와 그에 따른 사전 학습 및 파인튜닝 비용 절감(3.2절 참조) 덕분에, SLM은 에이전틱 시스템에서 등장할 때 대형 모델보다 본질적으로 더 유연하다. 따라서 다양한 에이전틱 루틴에 맞춘 여러 전문 모델을 훈련, 적응, 배포하는 것이 훨씬 더 저렴하고 실용적이 된다. 이 효율성은 빠른 반복과 적응을 가능하게 하여, 새로운 행동 지원, 출력 형식 요구 사항 충족, 특정 시장의 변화하는 지역 규제 준수 등 진화하는 사용자 요구를 충족할 수 있게 한다.

민주화: SLM이 LLM을 대체할 때 나타나는 결과 중 하나는 에이전트의 민주화이다. 더 많은 개인과 조직이 에이전틱 시스템 배포를 목표로 언어 모델 개발에 참여할 수 있게 되면, 전체 에이전트 집합은 보다 다양한 관점과 사회적 요구를 반영할 가능성이 높아진다. 이러한 다양성은 시스템적 편향 위험을 줄이고, 경쟁과 혁신을 촉진하는 데 도움을 줄 수 있다. 더 많은 참여자가 모델을 생성하고 개선함에 따라, 해당 분야의 발전 속도는 더 빨라질 것이다.

3.4 에이전트는 매우 제한적인 LM 기능만 노출한다

A4: 에이전틱 애플리케이션은 LM 기능의 제한된 하위 집합에 대한 인터페이스이다. 이는 V1과 V2 관점을 지지한다.

AI 에이전트는 강하게 지시받고 외부에서 조율된 언어 모델의 게이트웨이로, 본질적으로 사람-컴퓨터 인터페이스와 도구 선택 기능을 갖추어서, 올바르게 활용될 경우 유용한 작업을 수행한다. 이러한 관점에서, 강력한 범용성을 갖도록 설계된 대규모 언어 모델은 지루하게 작성된 프롬프트와 정교하게 조율된 컨텍스트 관리를 통해, 본래 가진 방대한 기술 범위 중 일부 제한된 영역에서만 동작하도록 조정된다. 따라서, 선택된 프롬프트에 적절히 파인튜닝된 SLM이면 충분하며, 앞서 언급한 효율성 향상과 유연성 증가라는 장점도 누릴 수 있다.

반론으로, 범용 LLM과의 신중한 인터페이스가 좁은 과제에서 강력한 성능을 위해 필요하다

고 주장할 수 있는데, 이는 LLM이 보다 넓은 언어 및 세계에 대한 이해를 갖고 있기 때문이라는 논리이다.

3.5 에이전틱 상호작용에는 긴밀한 행동 정렬이 필요하다

A5: 에이전틱 상호작용에는 긴밀한 행동 정렬이 필요하다. 이는 V2 관점과 일치한다.

일반적인 AI 에이전트는 LM 도구 호출을 통한 코드 상호작용이나, LM 호출을 수행하는 에이전틱 코드가 파싱할 출력 반환 등 코드와의 빈번한 상호작용을 한다. 이러한 상호작용의 성공을 위해서는, 생성된 도구 호출과 출력이 각각 도구 파라미터의 순서, 타입, 성격 및 LM을 호출하는 코드의 기대에 의해 부과된 엄격한 형식 요구사항을 준수해야 한다. 이런 경우, 모델이 여러 가지 다른 형식(JSON/XML/Python 도구 호출, XML/YAML/Markdown/Latex 출력)을 처리할 필요는 없으며, 에이전틱 애플리케이션 전체에서 일관성을 위해 하나의 형식만 선택하면 된다. 또한 모델이 때때로 환각 오류를 내며, 에이전틱 시스템의 "코드 부분"이 기대하는 형식과 다른 방식으로 응답하는 것은 바람직하지 않다. 이 때문에, 훈련 후 단일 형식 결정을 강제하거나 저비용의 추가 파인튜닝을 통해 장려된(encouraged) SLM이 AI 에이전트의 맥락에서는 범용 LLM보다 더 적합하다.

3.6 에이전틱 시스템은 본질적으로 이질적이다

A6: 에이전틱 시스템은 사용되는 모델 선택에서 본질적으로 이질성(heterogeneity)을 허용한다. 이는 V2 관점과 일치한다.

언어 모델 자체가 다른 언어 모델이 호출하는 도구(tool)가 될 수 있다. 마찬가지로, 에이전트의 코드가 언어 모델을 호출할 때마다 원칙적으로 어떤 언어 모델이든 선택할 수 있다. 이는 그림 1에서 보여진다. 복잡도가 다른 쿼리나 연산에 대해 크기와 능력이 다양한 여러 언어 모델을 통합하는 것이 SLM 도입을 위한 자연스러운 방법이다. 그림 1-왼쪽의 경우, 루트 에이전시를 가진 모델에는 LLM을 사용하고, 하위 LM에는 SLM을 사용할 수 있다. 그림 1-오른쪽에서는, 모든 LM이 원칙적으로 전문화된 SLM이 될 수 있다. 하나는 대화용으로, 다른 하나는 컨트롤러 정의 언어 모델링 작업 수행용으로 사용하는 식이다.

3.7 에이전틱 상호작용은 향후 개선을 위한 데이터 수집의 자연스러운 경로이다

A7: 에이전틱 상호작용은 향후 모델 개선을 위한 데이터의 좋은 원천이다. 이는 근본적으로 V2 관점을 지지한다.

3.4절에서 언급했듯, 에이전틱 과정에서 도구와 언어 모델을 호출할 때, 해당 시점에 필요한 제한된 기능을 제공하도록 언어 모델에 집중시키는 신중한 프롬프트가 종종 동반된다. 이러한 호출 하나하나가 향후 개선을 위한 자연스러운 데이터 원천이 된다(단, 처리되는 데이터가 보존 불가한 기밀 데이터가 아니라는 전제하에). 도구/모델 호출 인터페이스를 장식하는 리스너(listener)는 이후 전문화된 SLM을 파인튜닝하는 데 사용할 수 있는 특화된 지시 데이터를 수집할 수 있으며, 이는 향후 해당 호출의 비용을 줄이는 데 기여한다(그림 1의 로거 참조). 이 경로가 에이전트 아키텍처에 의해 가능하며, 높은 품질의 유기적 데이터를 생성한다고 주장한다(워크플로우 전체 성공률을 고려해 추가 필터링 가능). 따라서, 전문 SLM을 LLM 대신 배치하는 것은 에이전트 운영에서 단순한 부수적 노력이 아니라 자연스러운 단계가 된다.

5 채택 장벽

만약 A1-A7의 논거들이 정말로 설득력이 있다면, 왜 새로 등장하는 세대의 에이전트들은 여

전히 범용 LLM 사용이라는 현상 유지를 반복하는 것일까? SLM의 광범위한 채택을 가로막는 주요 장벽들이 다음과 같다고 본다:

B1. 중앙집중식 LLM 추론 인프라에 대한 대규모 초기 투자: 1장에서 자세히 설명했듯이, 미래 AI 서비스 제공의 주도적 패러다임이 중앙집중식 LLM 추론이 될 것이라는 전제 아래 막대한 자본 투자가 이루어졌다. 그 결과 업계는 이에 맞춘 도구와 인프라를 빠르게 구축해 왔으며, 보다 분산화된 SLM이나 온디바이스(on-device) 추론이 가까운 장래에 충분히 실현 가능할 수 있다는 가능성은 사실상 고려에서 배제되었다.

B2. SLM 훈련, 설계, 평가에서 범용 벤치마크 사용: 많은 SLM 설계 및 개발 연구가 LLM 개발의 궤적을 그대로 따라가며 동일한 범용 벤치마크에 의존하고 있다. 만약 평가 기준을 단순히 에이전트의 행위적 효용을 측정하는 벤치마크에 한정한다면, 연구된 SLM들이 대형 모델보다 손쉽게 더 뛰어난 성능을 보일 것이다.

B3. 대중적 인식 부족: SLM은 많은 산업 현장에서 더 적합함에도 불구하고, LLM에 비해 마케팅 강도나 언론의 주목을 충분히 받지 못한다.

장벽 B1-B3가 에이전틱 AI 맥락에서 SLM 기술의 근본적인 결함이라기보다는 실질적인 장애물에 가깝다고 본다. Dynamo와 같은 고도화된 추론 스케줄링 시스템의 등장으로 장벽 B1은 단순한 관성의 효과로 축소되고 있다. 장벽 B2 역시 점차 연구 분야에서 인식되고 있으며, 장벽 B3 또한 에이전틱 응용에서 SLM 배치가 가져올 경제적 이점(논거 A2)이 더 널리 알려지면 자연스럽게 해소될 것이다. 특히 장벽 B1이 지닌 관성을 고려할 때, 이러한 장벽들이 약화되거나 SLM이 대중적으로 채택되는 시점에 대해 구체적인 시한을 제시하지 않는다.

Class Imbalance 데이터의 학습 방법

취지: 기계학습 모델들은, 교사학습(Supervised Learning)의 대상이 되는 데이터들의 분포가 균형을 이루고 있어서 데이터 분포가 학습에 크게 영향을 미치지 않는다는 가정 하에, 다양하게 개발되었다. 그렇지만. 많은 분류 문제나 회귀 문제들은 데이터가 심각하게 분포가 한쪽으로 치우쳐있는 데이터 불균형을 다루고 있으며, 이 경우 일반적인 기계학습 모델들은 데이터 불균형 문제에서 제 성능을 발휘하지 못한다. 특히, 분류의 문제에서 소수 데이터를 지닌 영역의 class는 중요도가 아주 높기에 데이터 불균형에 대한 학습은 더 심각하게 다루어지고 있다. 따라서, class imbalance를 지닌 데이터를 학습하는 방법이 어떤 방식으로 다루어지고 있는지 알아본다. 그리고, 회귀에서 데이터 불균형이 일어나는 경우의 대처 방식도 알아본다.

조사문헌:

- S. Rezvani and X. Wang, “A Broad Review on Class Imbalance Learning Techniques,” Applied Soft Computing, vol. 143, 2023.

- N. Chawla et al., “SMOTE: Synthetic Minority Over-sampling Technoques,” J. Artificial Intelligence Reserach, vol. 16, 2002.

요약: 불균형 학습 문제는 비대칭적인 클래스 분포에서 이루어지는 학습 알고리즘의 성능에 영향을 미친다. 불균형 데이터셋은 복잡한 특성을 지니고 있기 때문에, 이러한 데이터로부터 효과적으로 학습하기 위해서는 새로운 알고리즘과 기존의 대용량 raw 데이터를 적절한 학습 데이터셋으로 변환할 수 있는 대처가 필요하다.

여기서는 데이터(클래스) 불균형 학습과 관련된 문제를 다루는 기존 기법들에 대하여 분류(classification) 작업과 회귀(regression) 작업을 대상으로 포괄적으로 살펴보았다. 또한, 불균형 학습 기법에 대한 분류 체계를 세 가지 범주로 나누어 정리하였다: (1) 데이터 전처리(Data Pre-Processing), (2) 알고리즘 구조(Algorithmic Structures), (3) 하이브리드 기법(Hybrid Techniques).

각 기법의 장점과 단점을 논의하고, 불균형 데이터셋의 분포에서 발생하는 주요 어려움을 설명하며, 이러한 문제를 해결하기 위해 제안된 주요 접근 방법들을 다루어본다. 마지막으로, 향후 이 분야의 연구를 촉진하기 위해 불균형 데이터를 활용한 학습 알고리즘 연구에서 연구 방향 설정에 도움이 될 수 있는 도전 과제를 제시한다.

1. 서 론

불균형 데이터셋의 분류(classification)는 기계학습 기술이 다루는 주요 문제 중 하나이다. 불균형 데이터셋이란, 한 클래스에 속하는 데이터 샘플의 수가 다른 클래스에 속하는 데이터 샘플 수보다 훨씬 많은 경우를 의미한다. 이 경우, 샘플 수가 많은 클래스를 다수 클래스(majority class), 샘플 수가 적은 클래스를 소수 클래스(minority class)라고 한다. 그림 1은 2차원 공간에서의 예시로, 소수 클래스(파란색 클래스)의 샘플 수는 다수 클래스(빨간색 클래스)보다 훨씬 적다. 보안, 금융, 인터넷 등과 같은 복잡하고 대규모의 네트워크로 이루어진 시스템에서 데이터 가용성이 지속적으로 증가함에 따라, 이 분야에 대한 기초적인 지식과 이해가 점점 중요해지고 있다. 기존 기계학습 기술들이 많은 실제 응용 분야에서 큰 성공을 거두었지만, 불균형 데이터에서의 학습 문제는 비교적 새로운 이슈이다.

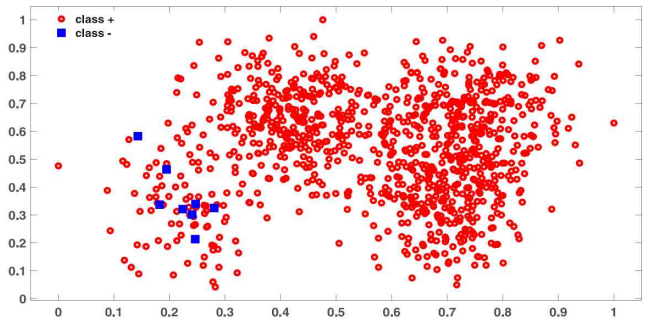


Fig. 1. Imbalanced dataset.

그림 2는 2009년부터 2021년까지 Information Science, Knowledge-Based Systems, Neural Networks, Applied Soft Computing, Expert Systems with Applications, Neurocomputing 여섯 개 저널에서 불균형 데이터의 학습을 주제로 발표된 논문의 추세인데, 불균형 데이터 학습 분야에서 발표된 논문의 수가 눈에 띄게 증가하고 있다. 이 분야의 빠른 발전에 따라, 과거와 현재의 데이터 불균형 문제의 학습에 관련된 연구를 유연하게 평가해보는 것이 향후 연구를 위해 필수적이다.

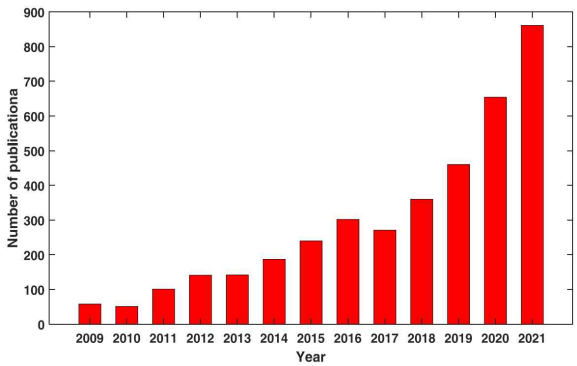


Fig. 2. Published papers on imbalanced domains.

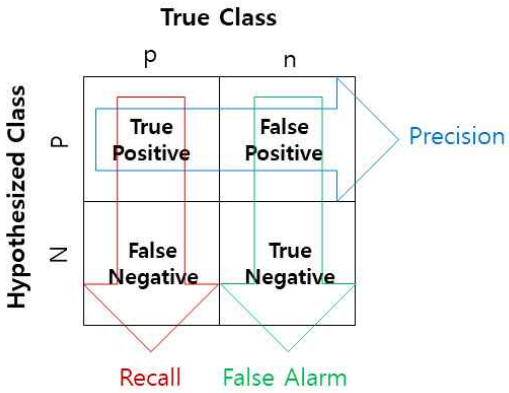
불균형 분류 문제를 처리하기 위해 다양한 SVM 변형 기법들이 있어 왔다. 불균형 분류 문제에 관한 여러 리뷰 논문들이 존재하지만, SVM을 중심으로 불균형 분류 문제를 심층적으로 다룬 리뷰는 아직 없는 실정이므로, 여기서는 SVM을 중심으로 불균형 학습 기법들에 대한 포괄적인 조사를 하고, 각 기법의 장점과 단점을 알아본다. 불균형 데이터셋 학습의 주요 어려움과 기회 역시 다루었다. 주요 사항은 다음과 같다.

(1) 분류와 회귀 작업에서 데이터 불균형 문제를 처리하는 방법들에 대해 설명한다.

(2) 클래스 불균형 학습 기법에 대한 분류 체계(taxonomy)를 제안한다.

(3) 각 유형의 불균형 학습 기법에 대한 장단점을 논의한다.
(4) 불균형 학습 분야에서의 주요 기회와 도전 과제를 강조하여, 향후 학습 알고리즘 연구 방향에 도움을 준다.

2. 데이터 불균형 학습의 성능 평가



클래스가 두 개인 경우 성능 지표의 기반 자료로 위 그림처럼 혼동행렬(Confusion Matrix)을 생성한다. 대표적인 성능 지표인 예측 정확도(predictive accuracy)는 클래스 불균형 데이터셋에서 성능을 평가하기에 부적절한 지표이다. 예를 들어, 다수 클래스가 99%, 소수 클래스가 1%의 비율을 갖는다고 가정하고, 만약 소수 클래스의 샘플이 하나도 올바르게 분류되지 않더라도, 예측 정확도는 99%이다. 하지만 불균형 데이터셋에서의 주요 과제는 소수 클래스 샘플을 정확히 탐지하는 것인데, 소수 클래스는 정확도가 0%이다.

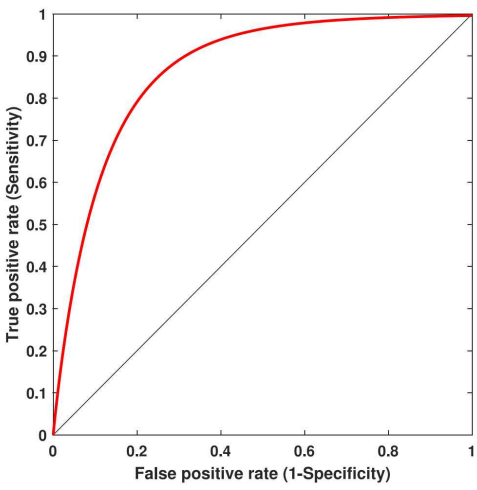


Fig. 4. ROC curve.

성능 평가의 대안으로 사용하는 방법은 특이도(Specificity)와 민감도(Sensivity)-상기율(Recall)-의 기하평균을 사용하는 방법이다. 이 지표는 p와n 클래스의 데이터 수 불균형에 영향을 받지 않는다.

$$Specificity = \frac{TN}{TN+FP}, \quad Sensitivity = \frac{TP}{FN+TP},$$

$$G-Mean = \sqrt{Specificity \times Sensitivity}$$

ROC(Receiver Operating Characteristic)는 신호검출기에서 검출 확률(Detection Probability)과 오류경보율(False Alarm Rate) 사이의 이율 배반성(Trade-Off)를 나타내기 위하여 사용된 방법인데, 이 방법이 기계학습 분야에서 분류기의 성능을 보여주고 비교하기 위하여 채택되었다. 이 방법은 TPR(=Sensitivity)과 FPR을 측정하여, 2차원 평면에 Fig. 4와 같이 그래프를 그리는데, p와 n 클래스의 데이터 수 불균형에 영향을 받지 않는다. ROC에서 AUC(Area Under Curve)를 측정하여 하나의 수치로 성능을 평가하는 것도 좋은 방법이다.

$$TPR = \frac{TP}{TP+FN} \qquad FPR = \frac{FP}{FP+TN}$$

4. 데이터 불균형 학습을 처리하기 위한 모델링 전략

불균형 도메인에서 예측 모델을 구축할 때 직면하게 되는 중대한 도전 과제를 해결하기 위해 사용되는 여러 전략들을 다음의 세 가지 주요 범주로 분류하였다:

(1) 데이터 전처리 (Data Pre-processing)

(2) 알고리즘 구조 (Algorithmic structures)

(3) 하이브리드 기법 (Hybrid techniques)

4.1. 데이터 전처리

데이터 전처리 기법은 주어진 불균형 데이터셋을 사전에 처리하여 데이터 분포를 수정함으로써 표준 알고리즘들을 효과적으로 사용할 수 있도록 하는 것이다. 즉, 학습 기법을 데이터셋에 직접 적용하기보다는, 먼저 데이터 분포를 수정하는 방식으로 전처리를 수행하는 것이다. 이러한 전처리 기법은 다음과 같은 장점이 있다:

(1) 기존의 어떤 학습 기법에도 적용할 수 있다.

(2) 데이터 분포가 연구자의 목표에 맞게 사전에 조정되었기 때문에, 해당 목표에 맞추어서 (편향되어) 선택된 모델이 가장 설득력이 높을 것이다.

한편 다음과 같은 단점도 있다:

(1) 데이터 분포를 손실 함수(loss function)와 연결하는 것은 어렵다. 왜냐하면 기존의 데이터 분포를 적절한 새로운 분포로 정확히 변환(mapping)하는 것이 쉽지 않기 때문이다.

(2) 초기 알고리즘은 사용자가 제공한 정보, 즉 어떤 클래스가 소수(양성) 클래스 인지를 기반으로 하기에, 어떤 관측값이 관련성 있는 데이터인지 혹은 어떤 것이 'normal' vs ‘abnormal’ 샘플인지 정의하기가 어렵다.

(3) 새로운 합성 샘플(synthetic sample)을 생성하는 것이 어렵다.

(4) 합성 샘플을 생성할 때, 각 샘플의 목표값(target value)은 두 샘플의 목표 변수 값의 가중 평균으로 계산되므로, 합성 샘플의 목표 변수값을 정확하게 결정하는 것은 어렵다.

4.1.1. 분류 문제의 데이터 전처리

(i) 랜덤 샘플링: 재샘플링(re-sampling) 기법들은 랜덤 오버샘플링/언더샘플링, 클러스터링 알고리즘, 데이터 클리닝 방식, 진화 알고리즘 등 다양한 전략을 기반으로 설계되어 있다.

언더샘플링과 오버샘플링은 데이터 분포를 조정하기 위해 사용되는 주요 재표본화 기법으로써, 각각 데이터셋 내 특정 클래스의 영향력을 줄이거나 높여, 최적의 오분류 비용(misclassification cost)에 접근하고자 한다. 다수 클래스의 샘플을 언더셈플링으로 제거하면 불균형을 해소할 수 있다. 한편, 과도한 오버샘플링이 과적합 가능성을 높여 성능에 부정적인 영향을 줄 수 있다. 오버샘플링/언더샘플링은 클래스 불균형 문제를 처리하기 위한 간단한 기법이지만, 오버샘플링은 추가적인 계산 비용이 많이 들며, 언더샘플링으로 얻는 성능에 비해 비효율적일 수 있다.

언더샘플링은 배깅(bagging)이나 BRF(Balanced Random Forest) 같은 앙상블 학습과 결합되어, 데이터셋의 분포를 조정하는 데 사용된다. 무작위로 샘플을 선택하는 것은 효과적인 전략일 수 있지만, 오버샘플링/언더샘플링 기법 그 자체가 강력한 기법으로 활용될 수 있다. 예를 들어, 언더샘플링은 거리 기반으로 수행될 수 있다. 이러한 기법들은 멀리 떨어진 예시

를 선택함으로써 양성 클래스를 지나치게 일반화하려는 경향이 있다. 또한, 대규모 데이터셋을 다룰 때 시간 소모가 많다.

불균형 데이터셋의 언더샘플링에서 분류 정확도와 데이터 분포의 균형을 맞추기 위해서는 진화 알고리즘(Evolutionary Algorithms)을 사용할 수 있다. 이 기법의 목적은 새로운 적합도 함수(fitness function)를 적용하여, 불균형 데이터셋 문제를 해결하는 진화 기반 프로토타입 선택 알고리즘을 제시하는 것이다.

(ii) 데이터 합성: 소수 클래스에서 새로운 데이터를 합성하는(synthesize) 방식으로 데이터 불균형을 해소한다. 새로운 데이터 합성은 여러 가지 이점을 가지고 있다, 예) 동일한 샘플이 학습 데이터에 반복해서 포함될 경우 발생하는 과적합 위험 축소. 오버샘플링으로 손상되기 쉬운 일반화 능력의 향상.

데이터 합성 기법은 크게 두 가지 그룹으로 나누어진다: 기존 샘플 간의 보간(interpolation) 방법, 잡음 추가 혹은 변형(perturbation) 방법.
보간을 사용하는 대표적인 기법은 **SMOTE**(Synthetic Minority Over-sampling Technique)이다. 이 방법은 다양한 불균형 비율과 학습 데이터 크기를 가진 여러 데이터셋에서 실험이 되었는데, 소수 클래스에 대한 오버샘플링과 다수 클래스에 대한 언더샘플링을 결합하여 적용한다.

그러나 SMOTE 알고리즘에는 다음과 같은 단점이 존재하여, 다양한 변형 기법들이 등장하게 되었다:
(1) SMOTE를 적용한 이후 또는 이전에 별도의 전처리나 후처리 과정이 필요함.
(2) SMOTE는 입력 공간의 특정 영역에서만 효과적으로 사용될 수 있음.

(iii) 적응형 합성 샘플링(Adaptive Synthetic Sampling): 다수 클래스는 그대로 유지하면서 소수 클래스에 대해 노이즈가 추가된 복제 샘플을 생성하는 적응형 합성 샘플링 방식이 제안되었다. 원래 데이터셋으로부터 노이즈가 추가된 여러 개의 학습 데이터셋을 만들고, 다양한 모델 추정값의 평균을 취함으로써 모델에 종속되지 않는 정규화(regularization) 방법을 제공할 수도 있다.

SMOTE을 다른 형태로 적용하여 특정 지역에서만 합성 샘플을 생성할 수 있다, 이 방식은 학습 알고리즘에 유용하지만, 이때 “좋은 영역(good region)”의 정의는 명확하지 않다. 이를 해결하기 위해 다양한 전략들이 제안되었다. 일부 방법은 클래스 간 경계 영역에 합성 노력을 집중하고, 또 다른 방법들은 학습이 어려운 샘플을 식별하고 해당 경계에 집중한다.

(iv) 데이터 정제 기법을 활용한 샘플링: 언더샘플링 기법은 데이터 정제(data cleaning) 기법을 통해서도 수행할 수 있다. 이러한 기법들의 주요 목적은 노이즈가 있는 샘플이나 클래스 간 겹치는 영역에 있는 샘플들을 식별하고 제거하는 것이다(A*방법). 이 기법들은 잘못된 경계 쪽에 있는 대부분의 샘플들을 제거함으로써 데이터셋의 크기는 줄어들고 품질은 향상된다.

(v) 클러스터 기반 샘플링 기법: 클러스터 기반 오버샘플링(CBO: Cluster-Based

Oversampling) 기법도 제안되었는데, 이 CBO 기법은 클래스 간 및 클래스 내 불균형을 동시에 처리할 수 있으며, 서브 클러스터(sub-cluster)에 대해 랜덤 오버샘플링을 수행한다. 그러나 이로 인해 오버피팅(overfitting) 문제가 발생할 수 있다. 클러스터링 기법에 대한 다른 아이디어도 여럿 있다.

(vi) 샘플링과 부스팅의 통합: SMOTE 기법은 부스팅 및 배깅과도 결합된다. 이 기법은 소수 클래스의 샘플에 대해 예제를 합성하며, 다수 클래스 샘플은 과도한 일반화(over-generalization)로 인해 무시된다. 하지만, 소수 클래스 샘플이 매우 분산되어 있고 클래스 불균형이 심할 경우, 이 기법은 특히 문제가 될 수 있다. 따라서 클래스 혼합(class mixture)을 사용하는 것이 더 나을 수 있다.

4.1.2. 회귀(regression)를 위한 데이터 전처리

분류 문제의 데이터 전처리 방법들은 다양하게 존재하지만, 회귀 문제에 대해서는 새로운 합성 데이터를 생성하기 위한 방법은 많지 않다. 회귀 작업의 경우, 희귀한(target이 드문) 경우와 가장 빈번한 경우 간의 불균형 문제를 극복하기 위해 데이터셋의 분포를 수정하는 방법이 제안되었다. 이 방법은 SMOTE 알고리즘을 수정하여 회귀 문제에도 적용할 수 있도록 하였다. 무작위 샘플 선택에 의존하는 기존의 방식과 달리, 오버샘플링/언더샘플링 접근법은 또 다른 지식 기반 기법으로 처리될 수 있다. 회귀 작업에 SMOTE를 적용하기 위하여 SMOTE을 수정한 기법인 SMOTER도 있다. 이는 SMOTE를 회귀 문제에 적용하여 희귀한 연속형 값의 예측 성능을 향상시키는 데 초점을 맞추었다.

회귀 문제에 SMOTE를 적용하기 위해 해결해야 할 세 가지 주요 질문은 다음과 같다:
(1) 어떤 관측값(observations)이 적절하고 일반적인 케이스인지 어떻게 설명할 수 있는가?
- 원래의 알고리즘은 학습자가 제공한 정보, 즉 어떤 클래스 값이 목표 클래스인지에 대한 정보를 기반으로 적용된다.
(2) 언더샘플링 또는 오버샘플링처럼 새로운 합성 샘플을 어떻게 생성할 수 있는가?
- 새로운 샘플을 생성하기 위해, 기존 SMOTE 알고리즘과 유사한 접근 방식이 사용되는데, 명목형 속성(nominal attribute)과 수치형 속성(numeric attribute) 모두를 처리할 수 있도록 몇 가지 소폭의 수정이 도입되었다.
(3) 새롭게 생성된 합성 샘플의 목표 변수 값을 어떻게 선택할 수 있는가?
- 생성된 합성 샘플들의 목표 변수 값을 어떻게 설정할 것인가에 관련된 질문이다.

4.2 알고리즘 구조(Algorithmic Structure)

알고리즘 구조 접근 방식은 불균형 문제를 처리하기 위해 새로운 알고리즘을 만들거나 기존 알고리즘을 수정하는 방법이다. 이러한 방법들은 다음과 같은 장점이 있다:
(1) 학습 목표가 모델에 직접적으로 반영된다.
(2) 모델을 이해하기 훨씬 더 쉽다.
하지만, 다음과 같은 단점도 존재한다:
(1) 비용 또는 비용-편익(cost/cost-benefit) 행렬이 제공되지 않는 경우가 많다.
(2) 데이터 분포에 맞게 적절한 알고리즘을 선택하기 위한 지식이 필요하다.
(3) 기존 학습 방법과 함께 사용하기 어렵다는 점에서, 전처리 기법과는 대조된다.

4.2.1 분류를 위한 알고리즘 구조

(i) 비용 민감 학습(Cost-Sensitive Learning): 샘플에 가중치를 부여하는 방식으로 비용 민감 학습(cost-sensitive learning)을 구현할 수 있다. 데이터셋 내에서 오분류 비용(misclassification cost)을 고려하여 더 나은 학습 분포를 선택할 수도 있다. 많은 경우, 소수 클래스의 일부 샘플을 식별하지 못하는 비용이 매우 크므로, 오분류 비용을 고려하지 않는 분류기는 성능이 저하되고 모델이 거의 쓸모없게 된다.

(ii) 데이터 공간 가중치 부여와 적응형 부스팅: 부스팅(boosting)은 여러 개의 약하고 정밀하지 않은 규칙들을 조합하여 높은 예측 정확도를 만들어내는 방법이다. AdaBoost는 여러 예측기를 생성하고, 약한 분류기들 사이에서 가중치 기반 투표(weighted voting)를 수행하는 앙상블 기법이다. 이 기법은 모든 오분류된 샘플에 동일한 가중치를 부여하지만, 실제로는 각 클래스마다 오분류율이 다르다. 보통, 소수 클래스의 오분류율이 다수 클래스보다 높으므로, AdaBoost 알고리즘은 데이터 분포가 불균형(skewed distribution)을 이루는 상황에서 작은 마진(small margin)과 높은 편향(high bias)을 초래한다. AdaBoost를 변형하여 가중치 갱신(weight update) 공식에 비용(cost)을 도입하고, 이 과정에서 사용된 가중치 갱신 규칙은 각각 다르게 설계할 수도 있다.

(iv) 신경망 기반 알고리즘: 신경망 기반의 다양한 분류기를 목적 함수의 수정으로 비용 민감형(cost-sensitive)으로 변형시켜 불균형 데이터 학습을 수행할 수 있다. 즉, 다수 클래스는 학습 강도를 낮추고 소수 클래스는 학습 강도를 높여준다.

또한, 불균형 클래스에 대한 보상을 위해 훈련 과정에 비용 함수(cost function)를 도입하고, 데이터 확률 분포에서 이 비용 함수의 영향을 줄이기 위한 전략도 제안되었다. 마지막으로, 앙상블 학습(Ensemble Learning) 역시 불균형 문제를 해결하기 위한 비용 민감형 프레임워크 내에서 고려되고 있다.

4.3. 하이브리드 기법

클래스 불균형 문제를 해결하기 위한 하이브리드 기법은 두 가지 이상의 방법으로 구성되며, 불균형 문제를 처리하거나 전체 솔루션의 특정 부분에 여러 접근 방식을 사용한다. 이러한 기법은 서로 다른 방법들이 적절히 상호 보완됨을 바탕으로 한다. 하이브리드 기법은 다양한 유형의 전략을 혼합하여 각 방법의 장점을 최대한 활용하려고 시도한다. 이러한 접근 방식의 주요 단점은 다음과 같다:
(1) 완벽하게 균형을 이루는 데이터가 반드시 최적이라고 할 수 없다.
(2) 이 시스템에서 적절한 오버샘플링/언더샘플링 비율을 결정하기가 어렵다.

4.3.1. 분류를 위한 하이브리드 기법

(i) 재샘플링과 학습 방법의 결합: 기계학습에서 앙상블은 여러 개의 분류기를 혼합함으로써 개별 분류기의 정확도를 향상시키는 것으로 잘 알려져 있다. 하지만 이들 앙상블 학습만으로는 클래스 불균형 문제를 단독으로 해결할 수 없다. 이러한 문제를 해결하기 위한 앙상블 학습 알고리즘 설계가 필요하다.

초기 하이브리드 기법은 어떤 분류기를 혼합할지, 그리고 어떤 방식으로 혼합할 수 있는지를 찾고자 하였다. 전문가 혼합(mixture-of-experts) 방식은 여러 재샘플링 기법을 전문가 프레임워크와 결합하는 것이 튜닝 문제에서 효과적이라고 가정한다. 신규 하이브리드 기법

중 하나는 부스팅과 데이터 샘플링을 결합한 것으로, RUSBoost라 불린다. 이 기법은 다음 두 가지 주요 재샘플링 방식을 다룬다:
(1) 소수 클래스 샘플을 다수 클래스 샘플과 거의 같은 크기가 되도록 오버샘플링
(2) 다수 클래스 샘플을 소수 클래스와 유사한 크기로 줄이는 언더샘플링
이 방법은 전문가 프레임워크 내에서 다양한 비율로 언더/오버샘플링을 결합하여 여러 분류기를 혼합하는 기법을 제안했다. 또한, 언더샘플링이 오버샘플링보다 더 효율적일 수 있다거나 어느 정도의 샘플링 비율이 적절한지에 대해서는 알려지지 않았음을 지적하였다.

다중 클래스 불균형 데이터를 대상으로 하는 앙상블 학습 기법이 제안되었다. 이 기법은 훈련 샘플의 왜곡된 정규분포 하에서 다중 클래스 샘플을 학습하는 일대다(one-versus-others) 분류기를 생성한다. 또 다른 앙상블 학습 기법인 eKISS (ensemble knowledge for imbalance sample sets)는 기본 분류기의 규칙을 결합하여 최종 의사결정을 위한 새로운 분류기를 생성하는 방식이다. 보다 복잡한 기술이 사용된 불균형 데이터를 위한 동적 분류기 앙상블 기법(DCEID, a dynamic classifier ensemble technique for imbalanced data)도 제안되었다. 이 기법은 앙상블과 비용 민감 학습을 결합하여 불균형 데이터셋에 대한 동적 분류기 앙상블 방식을 도입하였다.

비록 반례(counterexamples)를 사용하여 분류 경계를 더 정확하게 조정할 수 있는 다양한 형태의 일 클래스(one-class) 분류기가 존재하지만, OCC(One-class classification)는 모든 객체 중 특정 클래스의 객체만을 식별하고자 한다. 이 기법은 전통적인 분류 문제보다 훨씬 더 복잡하며, 모든 클래스 샘플이 포함된 훈련 세트를 바탕으로 두 개 혹은 그 이상의 클래스를 구분하려고 한다.

4.3.2. 회귀를 위한 하이브리드 기법

일반적으로 단일 계층의 로지스틱 회귀에 혼합 기법을 사용한다. 문제에서 최상의 결과를 달성하기 위해 이 아이디어를 효율적으로 활용하여 결과 집합을 미세 조정한다. 이 방법의 주요 장점은 여러 기법을 사용할 수 있다는 점이다. 따라서 가능한 모든 기법의 유용한 요소들을 통합하여 제공할 수 있다. 하지만 확장성(scalability)에는 문제가 있을 수 있으며, 그럼에도 불구하고 새로운 모델의 정확도는 큰 향상을 보여준다.

7. 향후 연구

불균형 데이터 영역에서 향후 연구 방향을 위한 여러 질문을 제시한다. 이 질문들은 불균형 데이터셋의 어려움을 해결하기 위한 다양한 기법들에서 직접적으로 도출된 것이다. 다음과 같은 근본적인 질문들이 불균형 학습 문제의 본질을 이해하기 위해 실험적 및 이론적 측면 모두에서 강하게 고려되어야 한다.

7.1. 클래스 불균형 학습에서의 데이터 전처리 기법

데이터 전처리 기법은 주어진 불균형 데이터셋을 전처리하고, 데이터 분포를 조정하여 사용자가 일반적인 알고리즘을 사용할 수 있도록 하는 전략들로 구성되어 있다. 다음의 질문들이 매우 신중히 연구되어야 한다고 생각한다:
(1) 데이터 분포와 목표 손실 함수(target loss function)를 연결하는 것은 어렵다. 이로 인해 데이터 분포를 최적의 새로운 분포로 매핑(mapping)하는 것이 어렵다. 따라서 데이터 분포를 원하는 새로운 분포로 어떻게 매핑할 것인가?

(2) 불균형 데이터 문제를 처리하기 위한 전처리 기법은 매우 다양하다. 하지만 모든 불균형 데이터셋에 적용 가능한 완벽한 기법을 찾는 것은 어렵다. 그래서 불균형 데이터셋을 다루기 위해 적절한 전처리 기법을 어떻게 식별할 것인가?
(3) 불균형 데이터를 처리하기 위한 전처리의 한 방법은 합성 데이터를 생성하는 것이다. 이는 실제 데이터만큼 좋은 데이터 생성으로 데이터 품질 향상에도 기여할 수 있다. 그러나 완벽한 합성 데이터를 생성하는 방법을 찾는 것은 어렵다. 따라서 어떻게 하면 완벽한 새로운 합성 샘플을 생성할 수 있을까?
(4) 새로운 합성 샘플을 생성할 때, 각 샘플의 목표값은 두 샘플의 목표 변수 값의 가중 평균으로 계산된다. 따라서 합성 샘플의 목표 변수 값을 결정하는 것이 어렵다. 그래서 합성 샘플의 목표 변수 값을 어떻게 설정할 것인가?
(5) 대규모 데이터셋은 부정확한 예측 가능성을 줄여주지만, 동시에 데이터 내의 약한 패턴을 찾는 것을 어렵게 만든다. 그리고 대량의 데이터를 저장하고 처리하는 것은 지능형 알고리즘 분야에서 도전 과제이다. 대규모 데이터셋에서 계산 복잡성을 어떻게 처리할 것인가?
(6) 노이즈가 있는 데이터란 불필요하고 의미 없는 정보가 다수 포함된 데이터를 말하며, 적절히 처리하지 않으면 분석 결과에 부정적인 영향을 끼치고 결론을 왜곡할 수 있다. 따라서 노이즈가 많은 데이터셋을 처리하는 방법을 어떻게 찾을 수 있을까?

이러한 문제들을 다루기 위한 제안된 해결책은 다음과 같습니다:
(i) 데이터 정제(Data Cleaning): 전처리 단계에서 가장 중요한 측면은 잘못되거나 부정확한 관측값을 찾아 수정하여 데이터 품질을 향상시키는 것이다. 이 기술은 데이터 내의 누락된 값, 부정확한 값, 중복 값, 관련 없는 값 또는 null 값을 식별하는 데 중점을 둔다.
(ii) 딥 생성 모델(Deep Generative Models): VAE(Variational Auto-Encoder)나 GAN(Generative Adversarial Network)과 같은 딥러닝 기반 생성 모델은 합성 데이터를 생성할 수 있다. VAE는 비지도 학습 방식으로, 인코더가 원본 데이터를 더 압축된 구조로 변환하여 디코더에 전달하고, 디코더는 이를 통해 원본 데이터를 나타내는 출력을 생성한다. 이 시스템은 입력과 출력 데이터 간의 상관관계를 최적화하면서 훈련된다.
(iii) 계산 복잡성 완화를 위한 속도 향상 기법: 계산 복잡성을 처리하기 위한 한 가지 방법은 좌표 하강법(Coordinate Descent)이다. 이는 함수의 최소값을 찾기 위해 각 좌표 방향을 따라 순차적으로 최소화를 수행하는 최적화 알고리즘이다. 각 반복에서 알고리즘은 선택 규칙에 따라 좌표 또는 좌표 블록을 선택한 뒤, 나머지 좌표를 고정한 채 해당 좌표 평면 위에서 함수 값을 최소화한다. 현재 위치에서 좌표 방향을 따라 선형 탐색을 수행하여 적절한 이동 크기를 결정할 수 있다.
(iv) 노이즈 감소를 위한 퍼지 집합(Fuzzy Set): 노이즈의 부정적인 영향을 줄이기 위한 효과적인 기법 중 하나는 퍼지 집합 이론이다. 예를 들어 퍼지 SVM(Fuzzy SVM)에서는 각 샘플에 퍼지 멤버십 값을 부여하고, 서로 다른 입력 샘플이 결정 경계 학습에 서로 다른 기여를 하도록 SVM을 재정의한다. 이 기법은 이상값 및 노이즈가 많은 데이터 포인트로 인한 부정적 영향을 줄이는 데 도움이 되며, 모델링되지 않은 특성을 가진 데이터 포인트가 포함된 응용 분야에도 적합하다.

7.2. 클래스 불균형 학습에서의 알고리즘 기법

불균형 영역에서의 알고리즘 접근 방식에서 향후 연구를 위한 질문은 다음과 같다:
(1) 노이즈가 있는 데이터셋을 처리하기 위한 다양한 전략의 여러 기법들이 존재한다. 하지

만 최적의 기법을 찾는 것은 매우 어렵다. 노이즈가 있는 데이터셋을 처리할 수 있는 기법들을 어떻게 비교할 수 있을까?
(2) 다중 클래스 분류 문제에서 불균형 데이터셋은 이진 분류 문제와는 다른 도전 과제를 지닌다. 알고리즘 구조 기법에서 최적의 다중 클래스 방법을 찾는 것은 어렵다. 다중 라벨 불균형 데이터셋에서의 분류 문제는 어떻게 처리할 것인가?
(3) 비용 민감 학습(Cost-sensitive learning)은 모델을 훈련시킬 때 예측 오류에 대한 비용을 고려한다. 이 접근 방식은 불균형 분류 문제에도 활용될 수 있다. 불균형 문제를 처리하기 위한 비용 민감 방법을 어떻게 찾을 것인가?
(4) 기존 기법을 다른 학습 시스템에 어떻게 적용할 수 있을까?
(5) 목표 손실 함수(target loss function)가 바뀌었을 때 모델을 어떻게 다룰 것인가? 모델을 다시 학습시켜야 할까, 아니면 추가적인 수정이 필요한가?
(6) 대규모 데이터셋에서의 계산 복잡성 문제는 어떻게 처리할 것인가?
위 문제들을 해결하기 위한 몇 가지 제안은 다음과 같다:
(i) 이진 학습기에서 다중 클래스 분류기로 전환: 다중 클래스 분류 문제에서는 One-vs-All(OVA, 일대다) 방식이 상대적으로 간단한 분해 전략으로 사용될 수 있다. 이는 각 클래스마다 해당 클래스와 나머지를 구분하는 이진 분류기를 학습시키는 방법이다.
(ii) 전처리 기법과 마찬가지로 계산 복잡성 처리에는 속도 향상 기법이 유효: 좌표강하법(Coordinate Descent)과 경사 하강법(Gradient Descent)은 계산 복잡성을 처리하기에 적합한 알고리즘이다.
(iii) 불균형 분류를 위한 비용 민감 학습: 비용 민감 분류는 잘못된 분류에 대해 서로 다른 비용을 부여한다. 이 중에서도 비용 민감 결정 트리(Cost-sensitive decision trees)는 가장 많은 관심을 받아왔다.
(iv) 노이즈 라벨 데이터를 다루기 위한 손실 함수의 수정: 손실의 단순한 불편 추정량(unbiased estimator)을 고려하고, 노이즈 라벨이 있는 데이터 하에서 경험적 효용 극대화(empirical utility maximization)를 위한 성능 경계를 도출할 수 있다. 또한, 노이즈 라벨 하에서의 위험 최소화 문제를 가중 0-1 손실(weighted 0-1 loss)을 사용하는 분류 문제로 환원할 수 있다. 이는 강력한 효용 경계를 얻을 수 있는 단순한 가중 대체 손실 함수(surrogate loss)를 사용할 수 있다는 것이다.

7.3. 클래스 불균형 학습에서의 하이브리드 기법

불균형 학습에서의 향후 하이브리드 방법 연구를 위해 다음과 같은 질문들이 존재한다:
(1) 데이터셋을 수정하기 위해 적절한 언더/오버 샘플링의 수를 어떻게 결정할 것인가?
(2) 매개변수와 다양한 커널에 대해 효과적인 전략을 어떻게 추정할 수 있을까?
(3) 대규모 데이터셋에서의 계산 복잡성 문제는 어떻게 처리할 것인가?
(4) 노이즈가 포함된 데이터셋을 어떻게 처리할 것인가?
(5) 다중 라벨 데이터셋에서의 분류 문제는 어떻게 해결할 것인가?
이러한 문제들을 해결하기 위해 전처리(pre-processing) 및 알고리즘 구조(algorithmic structure)의 제안들을 사용할 수 있다. 위에서 언급한 모든 질문들이 이론적인 연구뿐만 아니라 실용적인 응용 측면에서도 매우 중요한 핵심 과제이다.

8. 결론

본 조사에서는 불균형 학습 분야의 몇 가지 기법들과 주요 문제점들에 대해 논의하였다. 불

균형 데이터 영역에서의 평가 및 예측 모델링을 위해 여러 가지 방법들이 고려되었으며, 회귀(regression)와 분류(classification) 작업을 대상으로 다루었다. 또한, 클래스 불균형 학습 기법에 대해 분류 체계(taxonomy)를 제안하였으며, 각 기법은 그것이 기반하고 있는 동일한 전략에 따라 설명될 수 있도록 하였다. 불균형 처리 기법들의 분류는 다음과 같다: 데이터 전처리 (Data Preprocessing), 알고리즘 구조 (Algorithmic Structures), 하이브리드 기법 (Hybrid Techniques)

마지막으로, 불균형 데이터 분포와 밀접하게 관련된 여러 문제점들을 설명하고, 그러한 문제들과 불균형 데이터셋 사이의 관계를 탐색하였다.

SMOTE: Synthetic Minority Over-sampling Technique

데이터 불균형 문제를 다룰 때 가장 많이 사용되는 기법인 SMOTE는 소수 클래스 데이터를 합성하는 방식으로 오버샘플링을 하고 다수 클래스 데이터는 언더 샘플링을 하는 방식이다.

일반적으로 다수 클래스의 언더샘플링이 소수 클래스의 오버 샘플링보다 좋은 결과를 보여왔다. SMOTE는 소수 클래스의 오버샘플링을 샘플의 복제로 하는 것이 아니라 합성을 하는 방식을 취하고, 다수 클래스의 언더샘플링과 접목하여 좋은 결과를 보여준다.

소수 클래스의 합성에 의한 오버 샘플링 알고리즘(SMOTE)의 Pseudo-code는 다음과 같다. 한마디로 소수 클래스 샘플을 인접 소수 클래스 샘플과 선형적으로 보간하는 방식으로 소수 클래스에 속한 데이터를 합성한다.

Algorithm *SMOTE*(T, N, k)
Input: Number of minority class samples T; Amount of SMOTE $N\%$; Number of nearest neighbors k
Output: $(N/100)$ * T synthetic minority class samples
1. (∗ *If N is less than 100%, randomize the minority class samples as only a random percent of them will be SMOTEd.* ∗)
2. **if** $N < 100$
3. **then** Randomize the T minority class samples
4. $T = (N/100) * T$
5. $N = 100$
6. **endif**
7. $N = (int)(N/100)$ (∗ *The amount of SMOTE is assumed to be in integral multiples of 100.* ∗)
8. k = Number of nearest neighbors
9. $numattrs$ = Number of attributes
10. $Sample[\,][\,]$: array for original minority class samples
11. $newindex$: keeps a count of number of synthetic samples generated, initialized to 0
12. $Synthetic[\,][\,]$: array for synthetic samples
(∗ *Compute k nearest neighbors for each minority class sample only.* ∗)
13. **for** $i \leftarrow 1$ **to** T
14. Compute k nearest neighbors for i, and save the indices in the $nnarray$
15. Populate(N, i, $nnarray$)
16. **endfor**

Populate(N, i, $nnarray$) (∗ *Function to generate the synthetic samples.* ∗)
17. **while** $N \neq 0$
18. Choose a random number between 1 and k, call it nn. This step chooses one of the k nearest neighbors of i.
19. **for** $attr \leftarrow 1$ **to** $numattrs$
20. Compute: $dif = Sample[nnarray[nn]][attr] - Sample[i][attr]$
21. Compute: gap = random number between 0 and 1
22. $Synthetic[newindex][attr] = Sample[i][attr] + gap * dif$
23. **endfor**
24. $newindex$++
25. $N = N - 1$
26. **endwhile**
27. **return** (∗ *End of Populate.* ∗)
End of Pseudo-Code.

대량언어모델(LLM)을 활용한 과학연구 조사: LLM4SR

취지: LLM의 비약적 발전은 과학연구의 수행 방식까지도 영향을 끼치고 있다. 이에 현재 LLM이 과학 연구에서 가설 수립, 실험 계획 및 구현, 과학 문서 작성, 동료 심사의 4단계에 걸쳐 어떻게 활용되고 있으며, 어떠한 문제점이 존재하고, 어떠한 방향으로 나아가고 있는지를 살펴보고자 한다.

조사 문헌: Z. Luo, Z. Yang, Z. Xu, W. Yang, and X. Du, "LLM4SR: A Survey on Large Language Models for Scientific Research," https://arxiv.org/abs/2501.04306

1. 서론

전통적인 과학 연구 과정은 배경 지식 습득, 가설 수립(제안), 실험 설계 및 수행, 데이터 수집 및 분석, 문서 작성 및 동료 심사로 이루어진다.

이러한 과학 연구 과정에 대한 자동화 시도는 수십 년에 걸쳐 있어 왔다. 최근 LLM(Large Language Model)의 비약적 발전은 대량의 데이터를 처리하여, 자연어 문장을 생성하고, 복잡한 결정에 대한 지원도 이루어지게 해주었다. 이제 LLM이 과학 연구의 수행과 문서 작성 및 평가(심사)에 걸쳐 혁명적 변화를 이끌어 갈 수 있음을 보여주고 있다.

여기서는 LLM이 어떻게 과학 연구 과정 전반에 걸쳐 활용되고 있는지를 다음과 같이 4 단계로 구분하여 방법론을 조사한다.

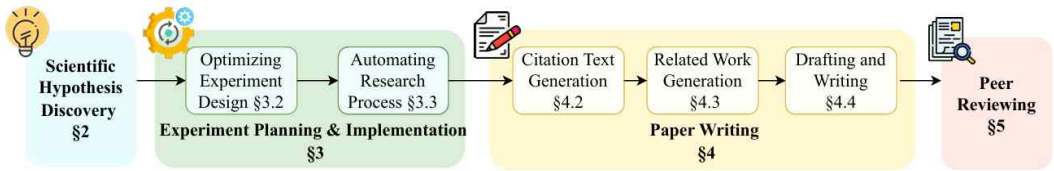
① **과학적 가설 수립(발견)(Scientific Hypothesis Discovery)**: 축적된 지식과 실험 데이터를 활용하여 고유한 과학 아이디어를 제안

② **실험 계획 및 구현(Experimental Planning and Implementation)**: 실험 설계를 최적화하고, 실험 과정을 자동화하며, 실험 데이터를 분석

③ **과학 문서 작성(Scientific Writing)**: 인용문헌 작성, 관련 연구 정리, 과학문서 초안 작성

④ **동료 심사(Peer Review)**: 논문 심사 자동화와 오류 확인 및 일관성 검토를 통한 논문 심사 지원

더 나아가, 각 업무의 한계와 향상이 필요한 부분을 명시한다. 그 결과, 연구자들로 하여금 새로운 개념을 탐구하고, 평가 방법을 개발하고, LLM을 효과적으로 업무에 활용할 혁신적 접근 방법을 구상하고자 하는데 도움을 주고자 한다.



2. 과학적 가설 수립

2.1. 개요

LLM을 활용한 과학 가설 발견의 출현 이전에 존재하였던 관련 연구의 주요 두 줄기인 문헌기반 발견(LBD: Literature-Based-Discovery)과 귀납추론(Inductive Reasoning)를 살펴보고, LLM을 이용한 방법론, 평가개발 트랜드, 중요한 진전, 주요 도전과제를 살펴본다.

2.2. 역사적 배경

2.2.1. 문헌기반 발견(LBD: Literature-Based-Discovery)

LDB의 핵심 개념은 다음과 같다: “지식은 공개된다. 그러나 독립적으로 생성된 지식 조각들이 논리적으로 관련되어 있지만 결코 검색되거나, 모아지거나, 해석되지 않았으면, 그 지식은 발견되지 않은 채로 남아 있을 수 있다”.

- 공개된 지식을 검색하여 새로운 지식을 창출하는 방법이 도전 과제이다.
- “가설은 축적된 지식의 연결과 협력으로 만들어질 수 있다”.

. Swanson: LBD의 고전적인 형식화를 제안했으며, 이를 "ABC" 모델이라고 한다. 이 모델에서는 개념 A와 개념 C가 특정 중간 개념 B와 논문에서 함께 등장하는 경우, A와 C가 연결될 가능성이 있다고 가정한다.

. Wang et al.: 기존 LBD 방법은 인간 과학자가 아이디어를 도출하는 과정에서 고려하는 맥락을 모델링하지 못하며, 개별 개념 간의 쌍(pairwise) 관계만을 예측하는 데 제한이 있다. 이러한 한계를 극복하기 위해 LBD를 자연어 맥락에 적용하여 생성 공간을 제한하는 첫 번째 시도를 수행했으며, 전통적인 LBD처럼 단순히 관계를 예측하는 것이 아니라 생성된 문장을 출력으로 사용하는 방식을 도입했다.

. Yang et al.: 각각 사회과학 및 화학 연구자들과 광범위한 논의를 진행한 결과, 기존에 발표된 대부분의 사회과학 및 화학 관련 가설(특정 유형의 가설에 한정되지 않음)이 LBD 패턴으로 공식화될 수 있음을 발견했다.

2.2.2. 귀납추론(Inductive Reasoning)

구체적인 관찰로부터 광의의 응용 관점을 도입하여 일반적인 법칙/규칙(rule)과 가설을 발견할 수 있다. 귀납추론에서 법칙/규칙(rule)이 갖추어야 할 기본 사항은 다음과 같다.

(1) “법칙/규칙(rule)”은 관찰과 충돌하지 않는다.
(2) “법칙/규칙(rule)”은 현상을 반영한다.
(3) “법칙/규칙(rule)”은 구체적인 관찰보다 더 넓은 영역에 적용되는 일반적인 패턴을 제시하며, 관찰에 나타나지 않는 새로운 정보를 지녔다.
(4) “법칙/규칙(rule)”은 명확하고 충분히 상세하여야 한다.

많은 예비 법칙/규칙을 생성한 후, 위의 4가지 기본 요구 사항에 의해 걸러낸다. 필터링을 대체하고 더 나은 규칙을 만들기 위해 더 많은 추론 단계를 사용하는 자체 정제(self-refine) 방법도 시도 되었지만, 이런 방식은 알려진 지식이거나 과학 지식이 아닌 경우도 있었다.

. Yang et al.: LLM을 활용하여 공개된 웹사이트의 정보를 기반으로 고유하고 유용한 사회과학적 가설을 수립하였다.

. Majumder et al., “데이터 기반 발견(data-driven discovery)”: 웹상의 모든 공개 실험 데이터(및 보유한 비공개 실험 데이터)를 활용하여 학문 간 가설을 발견한다. 이 방법은 대량의

공개 실험 데이터의 의미가 충분히 파악되지 않았으며, 기존 데이터로부터 많은 새로운 과학적 가설을 발견할 수 있다는 점에 근거한다.

2.3 Development of Methods (방법론 개발)

과학적 가설 수립의 방법론 중에 주요 흐름을 살펴보고, 다른 방법들도 알아본다.

2.3.1 주요 흐름(Main Trajectory)의 핵심 구성 요소(Key Components)

주요 흐름의 핵심 구성 요소는 다음과 같다. 이들을 각각 자세히 살펴본다.

- 영감 검색 전략(Strategy of Inspiration Retrieval)
- 피드백 모듈(Feedback Modules): 참신성 검사기(Novelty Checker), 타당성 검사기(Validity Checker), 명확성 검사기(Clarity Checker)
- 진화 알고리즘(Evolutionary Algorithm)
- 다중 영감 활용(Leverage of Multiple Inspiration)
- 가설 순위 결정(Ranking of Hypothesis)
- 자동 연구 질문 구성(Automatic Research Question Construction)

영감 검색 전략(Strategy of Inspiration Retrieval)

- LBD(Literature-Based-Discovery) 개념에 의한 방법

. SciMON: 기존 지식의 연결로 새 지식 구성. SentenceBERT로 임베딩하여 cosine similarity로 두 지식의 의미적 유사성(semantically similar)을 찾아낸다.

. ResearchAgent: "ABC" 모델을 엄격하게 따르며 개념 그래프를 구축하는데, 이 개념 그래프에서 배경 개념과 연결된 개념을 영감 개념으로 검색한다,

. Scideator: 의미적 매칭(semantic scholar API 추천)과 개념 매칭(같은 주제, 같은 하위 영역, 그리고 다른 하위 영역에서 유사한 개념을 포함하는 논문)을 기반으로 영감이 되는 논문을 검색한다.

. SciPIP: 의미적으로 유사한 지식(SentenceBERT 기반), 개념 공동 발생, 그리고 인용 그래프 이웃을 통해 영감을 검색한다. 또한, 개념 공동 발생 검색에서 유용하지 않은 개념을 걸러내는 필터링 방법을 제안한다.

. SciAgents: 의미적 또는 인용 관계 이웃을 영감으로 선택하는 기존 접근법과 달리, 배경 개념과 인용 그래프에서 (긴 경로나 짧은 경로를 통해) 연결된 다른 개념을 무작위로 샘플링하여 영감 요소로 사용한다.

- LLM 선택 영감(LLM-selected inspiration)

. MOOSE: 연구 배경과 문맥 내의 일부 영감 후보를 제공한 후, LLM에게 연구 배경에 적합한 영감을 후보들 중에서 선택 하도록 한다. 이 방법을 화학 분야에 적용한 MOOSE-Chem은 해당 가정이 상당 부분 타당할 수 있음을 보였다.

. NOVA: LLM이 선택한 영감을 채택한다. LLM의 내부 지식을 이용해 새로운 아이디어에 유용한 지식을 결정하는 것이 전통적인 엔티티/키워드 기반 검색 방법을 능가할 것이다.

Feedback Modules: 생성된 가설에 대한 참신성, 타당성, 명확성 관점의 반복적인 피드백이 핵심이다. 참신성 검사기(novelty checker), 타당성 검사기(validity checker), 명확성 검사기

(clarity checker)로 나누어 설명한다.

- **참신성 검사기(Novelty Checker)**: 생성된 가설이 참신성이 부족하면 feedback으로 참신성을 향상시키는 LLM 기반 방법이다.

. MOOSE: 생성된 가설을 관련 문헌 조사로 평가한다

. SciMON, SciAgents, Scideator, CoI: 관련 논문을 반복적으로 검색하여 비교한다.

. Qi, ResearchAgent, AIScientist, MOOSE-Chem, VirSci: LLM 내부 지식 직접 활용하여 평가한다.

- **타당성 검사기(Validity Checker)**: 생성된 가설은 객관적인 우주의 원리를 정확하게 반영하는 과학/공학적 발견이어야 한다. 가설의 실제 타당성을 평가하는 피드백은 실험 결과에서 나와야 하지만, 각 생성된 가설에 대해 실험을 수행하는 것은 시간과 비용이 많이 들기에 거의 전적으로 LLM 또는 기타 훈련된 신경망 모델의 휴리스틱(heuristics)을 사용한다.

- 이외에 다른 방법을 채택한 것은 다음과 같다.

. FunSearch: 수학 문제에 대한 code 생성

. HypoGeniC, LLM-SR: 생성된 각 가설에 대한 일관성을 유지하는 예제에 집중하는 방법

. SGA: 실제 실험을 모방한 가상 시뮬레이션 환경을 생성

타당성 검사는 여전히 아직 해결되지 않은 (가설 검증을 위한 자동 실험 수행 등이 필요한) 미래의 도전 과제이다.

- **명확성 검사기(Clarity Checker)**: 생성된 가설은 명확한 정보 전달과 적절한 세부 사항 제공이 되어야 하는데, LLM은 종종 세부 사항이 부족한 가설을 생성하는 경향이 있다. 가설을 정제하고 상세한 기술을 하는데 명확성 피드백이 유익하다.

. MOOSE, ResearchAgent, MOOSE-Chem, VirSci: LLM 활용 명확성에 대한 자체 평가를 수행한다.

진화 알고리즘(Evolutionary Algorithm)

환경에 적응하지 못하는 객체는 소멸되고, 환경에 적응하는 객체는 조합에 의해 진화(소위, “돌연변이”라고 지칭)되는 진화 알고리즘을 가설 수립에 적용한다.

(1) 생성된 가설의 실제 실험에 의한 평가와 휴리스틱 평가 모두 환경에 대입시킨다.

(2) 과학적 가설 발견의 핵심은 알려진 지식을 입력받아 아직 알려지지 않은 유용한 지식을 생성하는 돌연변이 생성으로 볼 수 있다.

. FunSearch: 각 island를 비슷한 방법의 그룹으로 구성하고, islands를 진화 알고리즘으로 개선시킨다. 순위가 낮은 islands는 제거되며, 재조합에 의해 좋은 islands로 구성된 새로운 islands를 생성한다.

. LLM-SR: FunsSearch와 유사한 아일랜드 기반 진화 알고리즘을 적용한다.

. SGA: 다중 자손을 생성하고 최고를 선택하는 방법이다. 과거 실험으로부터 더 좋은 탐색을 위해 교배에 의해 새로운 가설을 생성한다.

. MOOSE-Chem: 배경 지식과 영감 지식이 주어지면, 이 둘을 협력시켜 여러 개의 고유한 가설들을 생성하고, 각 가설은 독립적으로 정제된 후, 최종적으로 정제된 가설들을 다시 결합하여 배경 지식과 영감이 더 잘 통합된 응집력 있는 가설을 형성한다. 동일한 입력에서 다양한 돌연변이 변형(mutation variants)을 유도하고, 각 변형의 장점을 모아 통합시킨다.

다중 영감 활용(leverage of multiple inspiration)

여러 가지 영감을 명확하게 식별하여 이를 최종 가설에 모두 활용 한다. 방법이 다른 것은 다 이유가 있다는 것을 중시한다.

. MOOSE-Chem: 해결하기 어려운 문제 *P*(hypothesis|research background)를 보다 작고 실용적이며 실행 가능한 단계로 분해한다. 이 과정은 분해를 위한 수학적 증명을 공식화함으로써 이루어진다. 일반적으로 이 작은 단계들은 다음과 같은 과정을 포함한다.

*초기 영감을 식별하고,

*연구 배경과 영감을 바탕으로 예비 가설을 작성한 뒤,

*예비 가설의 부족한 부분을 보완할 추가적인 영감을 찾아내고,

*새로운 영감을 반영하여 가설을 업데이트하는 과정이 반복된다.

. Nova: 연속적인 방식으로 여러 영감을 검색하여 더 다양하고 참신한 연구 가설을 생성한다. (IGA의 실험 결과: 생성된 가설의 다양성이 일정 수준에서 포화되는 경향이 있다. 이러한 문제의 주요 원인 중 하나는 입력되는 배경 정보가 동일하다는 것이다.) 유연한 입력을 도입하여 다양한 영감들을 결합하도록 한다.

가설 순위 결정(ranking of hypothesis)

생성된 가설에 대한 전체 순위(full ranking)를 제공한다. LLM이 짧은 시간 내에 방대한 수의 가설을 생성할 수 있는 반면, 각 가설을 실험적으로 검증하기에는 긴 시간과 많은 비용이 필요하다. 따라서 과학자들이 어떤 가설을 우선적으로 검증해야 할지 아는 것은 매우 유익하다.

. 대부분의 방법(MCR, AIScientist, MOOSE-Chem, CycleResearcher): LLM이 평가한 점수를 보상 값으로 사용하여 순위를 결정해준다.

. FunSearch: 코드 생성 문제에 초점을 맞추고 있어, 생성된 코드를 실행하고 결과를 확인하는 방식으로 보다 정확하게 평가할 수 있다.

. ChemReasoner: 특정 작업을 위한 그래프 신경망(Graph Neural Network, GNN) 모델을 미세 조정하여 점수를 매긴다.

. HypoGeniC/LLM-SR: 데이터 기반 발견(data-driven discovery)에 초점을 맞추며, 가설과 일관성을 갖는 관찰 예제의 수를 점수로 활용하여 순위를 매긴다.

. IGA의 차별점: LLM이 최종 점수나 결정을 직접 예측할 때 정확도가 낮지만, 두 개의 논문을 비교하여 어느 것이 더 나은지를 판단하는 경우 상대적으로 높은 정확도를 보이기 때문에 쌍 비교 방법을 채택한다. CoI와 Nova도 동일한 방식의 쌍 비교 자동 평가 방법을 채택했다.

자동 연구 질문 구성(automatic research question construction)

연구 질문을 자동으로 구성하고 이를 입력 값으로 사용하여 가설을 발견할 수 있게 한다. 이 요소가 있다면, 시스템은 "완전 자율 주행(full-self-driving)" 모드로 작동하며, 인간의 입력 없이 독립적 발견을 수행 할 수 있다.

. MOOSE: LLM 기반의 에이전트를 채택하여 관련 분야의 웹 코퍼스를 지속적으로 검색하여 흥미로운 연구 질문을 찾는다.

. AIScientist: 시작 코드 구현을 입력으로 활용하여 연구 방향을 탐색한다.

. MLR-Copilot: 입력된 논문들을 분석하여 연구의 갭을 찾고, 이를 바탕으로 연구 방향을 제시한다.

. SciAgents/Scideator: 연구 질문 없이 개념들의 결합을 바탕으로 직접 가설을 생성한다.

. VirSci: LLM기반 과학자 에이전트를 활용하여 브레인스토밍을 통해 연구 질문을 생성한다.

. CoI: 방법론의 발전 방향을 수집하여 다음 단계를 예측함으로써 연구 질문을 찾는다.
. Nova: 연구 질문 구성 단계 없이, 입력된 논문과 일반적인 아이디어 제안 패턴에서 직접 아이디어를 생성한다.

2.3.2 다른 방법들

과학적 발견의 다양한 측면에 초점을 맞추고 있어 매우 다양한 방법들이 존재한다. 예를 들어, Dong et al.은 독특한 방법론을 활용하고, Pu et al.은 HCI(Human-Computer Interaction)에 집중하며, Liu et al.은 실험 결과의 통합을 고려한다. 또한, Li et al.과 Weng et al.은 리뷰를 선호도 학습으로 활용하여 가설 제안 모델을 정밀 조정하는 방법을 연구한다.

. Dong et al.: GPT-4를 활용하여 "P = NP인지 여부"라는 매우 어려운 연구 질문을 해결한다. “소크라테스적 추론(Socratic reasoning)”이라는 접근법을 제안하는데, 이는 LLM이 문제를 반복적으로 발견하고 해결하며 통합하도록 장려하는 동시에 자체 평가와 개선을 촉진한다. 이 방법은 기존의 매우 어려운 가설을 증명하려는 시도에 유용하다.
. IdeaSynth: 연구 아이디어 개발 시스템으로, 아이디어 개념을 캔버스의 연결된 노드로 표현한다. IdeaSynth를 사용한 참가자들은 강력한 LLM 기반 비교 모델을 사용한 참가자들보다 더 다양한 대안을 탐색하고 초기 아이디어를 더 세부적으로 확장할 수 있었다.
. Liu et al. “문헌 기반 발견과 데이터 기반 발견을 통합”: 초기 실험 결과가 주어지면, 관련 문헌을 검색하고 반복적인 개선 접근법을 채택하여 가설을 실험 결과와 일관되게 유지하면서 검색된 문헌에서 얻은 발견을 활용한다.
. Weng et al.: CycleResearcher와 CycleReviewer로 구성된 이중 시스템을 제안한다. CycleResearcher는 아이디어 구상과 논문 작성을 담당하고, CycleReviewer는 작성된 논문을 평가하는 역할을 한다. 이중 시스템은 CycleReviewer의 점수를 활용하여 CycleResearcher를 훈련할 수 있도록 함으로써 상호 시너지 효과를 얻는다. 이 시스템은 실험 계획 및 실행 단계를 생략하고 아이디어 구상과 논문 작성 기능만 수행한다.
. Li et al. “두 단계 접근법”: 대형 언어 모델(LLM)의 아이디어 생성 능력을 향상시키기 위해 미세 조정하는 방법을 제안하며, 지도 학습(SFT)과 제어 가능한 강화 학습(RL)을 결합한 두 단계 접근법을 도입한다. 이들은 실행 가능성(feasibility), 참신성(novelty), 효과성(effectiveness)이라는 세 가지 차원에 초점을 맞추며, 차원별 컨트롤러를 통해 생성 과정을 동적으로 조정할 수 있도록 한다.

2.5 평가 개발 추세

과학적 발견 과제(scientific discovery tasks)에 대한 평가는 아주 다양하다. 모든 논문이 다른 평가 방법론을 제시한다. 공통적 평가 지표로는 참신성(novelty), 타당성(validity) 혹은 실현가능성(feasibility), 명확성(clarity), 중요성(significance), 관련성(relatedness), 흥미성(interestingness), 유용성(helpfulness)이 있다.
타당성(validity)은 발견된 과학적 지식이 객관적 세계를 정확하게 반영하는지를 뜻하며, 실현가능성(feasibility)은 공학적 발견의 실용성을 나타낸다.

평가자에 근거한 구분: LLM 기반 평가와 전문가 기반 평가로 나누어진다.
LLM 기반 평가는 사회과학 분야에서 전문가 기반 평가와 높은 일관성을 보여주지만, 자연과

학 분야에서는 신뢰할 만하지 않다는 논쟁이 있다. 전문가 평가는 대체적으로 신뢰하지만, 화학에서 충분히 신뢰하지 못하는 경우도 있다. 그 이유 2가지는 다음과 같다.
(1) 해당 학문의 복잡성
(2) 연구 주제의 미세한 차이가 평가를 위해 완전히 다른 배경 지식을 요구
- 전문가들은 일반적으로 특정 연구 분야에 집중하는 경향이 있어, 신뢰할 만한 평가에 필요한 지식을 모두 갖추지 못할 가능성이 있다.

참고자료(레퍼런스) 활용 여부: 직접 평가(direct evaluation)와 참고 기반 평가(reference-based evaluation)로 구분될 수 있다.
직접 평가는 신뢰성 문제가 제기될 수 있기 때문에, 참고 기반 평가가 대안으로 활용될 수 있다. 참고 기반 평가는 생성된 가설에서 실제(ground truth) 가설에서 언급된 핵심 요소들이 생성된 가설에 얼마나 포함되었는지를 확인하는 방식이다.

또한, 생성된 가설에 단순히 점수를 부여하는 방식의 한계를 보완하기 위해, **비교 기반 평가**(comparison-based evaluation)가 제안되었다. 이는 LLM 평가자가 여러 가설을 계속 비교하며 상대적 순위를 매기는 방식으로, 두 가지 방법이 생성한 가설의 상대적인 품질을 평가하는 데 유용할 수 있다.

그렇지만, 최종 평가는 실제 실험(wet-lab)에 의한 것이어야 한다.

2.6 가설 발견의 주요 진전/성과

.Yang et al.: LLM이 새롭고 유효한 과학적 가설을 생성할 수 있음을 전문가 평가를 통해 처음으로 입증하였다. 그들은 세 명의 사회과학 박사 과정 학생들이 생성된 사회과학 가설의 새로움과 유효성을 직접 평가하도록 했다.
. Si et al.: 100명 이상의 NLP 연구자를 고용하여 LLM이 생성한 가설에 대한 대규모 전문가 평가를 처음으로 진행했다. 이들은 LLM이 인간 연구자들보다 더 새롭지만, 조금 덜 유효한 연구 가설을 생성할 수 있다는 통계적으로 유의미한 결론을 도출했다.
. Yang et al.: LLM 기반의 프레임워크가, 2023년 10월까지의 데이터만으로 학습된 LLM을 사용하여, 2024년에 Nature, Science 또는 그에 준하는 수준의 저널에 발표된 많은 화학 및 재료 과학 가설의 주요 혁신을 재발견할 수 있음을 보여주었다.

2.7 도전 및 향후 연구

도전 과제

과학적 발견은 실험실의 실험으로 검증되지 않은 새로운 지식을 찾는 과정이다.
- 이로 인해 대규모 머신 생성 가설을 검증하기 위한 자동화 실험 수행의 필요성이 제기됩니다.
- 현재의 과학적 발견 방법은 기존의 대형 언어 모델(LLM)의 능력에 크게 의존하고 있는데, 과학적 발견 작업에서 LLM의 능력을 어떻게 향상시킬 수 있는지에 대한 명확한 해답은 아직 부족하다.
- 과학적 발견을 위한 충분한 내부 추론 구조가 무엇인지 명확하지 않다. 현재 연구들은 고품질 지식 출처(예: 학술 문헌)를 검색하여 영감을 얻고 가설을 생성하는 방식에 의존하고 있

으나, 이 과정에서 활용될 수 있는 추가적인 내부 추론 구조가 존재하는지에 대한 연구는 부족하다.
- 정확하고 체계적인 벤치마크를 구축하는 과정에서 전문가의 역할이 필수적이다.

향후 연구 방향
- 자동화된 실험 실행을 강화한다.(가설의 유용성을 확인하기 위한 가장 신뢰할 만한 방법)
. 컴퓨터 분야: 코딩 능력
. 화학/생물학: 실험 수행 로봇
- LLM의 가설 생성 능력을 향상시킨다.(가능한 방법: 학습 데이터 수집과 학습 전략)
- 과학 발견 과정의 다른 내부 추론 구조(internal reasoning structure)를 탐색한다. 이 과정에서는 과학 철학(Science of Science) 등 학제 간 연구가 필요할 수 있다.
- LLM을 활용하여 정교하고 구조화된 벤치마크를 구축한다.

3. 실험 계획 및 구현을 위한 LLM
3.1 개요
- 과학 연구에서 실험 설계를 자동화하고 워크플로를 간소화하는 데 LLM을 활용한다.
- LLM은 방대한 내부 세계 지식을 보유하고 있어 특정 도메인 데이터를 학습하지 않고도 현실 세계에서 정보에 기반한 작업을 수행할 수 있다.
- 이러한 잠재력을 극대화하기 위해, LLM은 두 가지 핵심 속성인 모듈성(modularity)과 도구 통합(tool integration)을 갖춘 에이전트 기반(agent-based) 방식으로 설계된다.
- 모듈성: LLM이 데이터베이스, 실험 플랫폼, 계산 도구와 같은 외부 시스템과 원활하게 상호작용할 수 있도록 보장한다.
- 도구 증강(tool-augmented) 프레임워크: LLM이 데이터 검색, 계산, 실험 제어를 위한 특화 모듈과 인터페이스를 형성하여 워크플로의 중심 컨트롤러 역할을 수행하도록 한다.

3.2 실험 설계 최적화(Optimizing Experimental Design)
- LLM의 방대한 데이터셋을 처리하고 분석하는 능력을 통해 연구자들은 복잡한 작업을 세분화하고, 최적의 방법론을 선택하며, 실험의 전체적인 구조를 개선한다.

- **작업 세분화(Task Decomposition):** 실험을 보다 작고 관리 가능한 하위 작업으로 나누는 과정으로, 현실 세계 연구의 복잡성으로 인해 연구 목표와의 일치를 보장하기 위해 필수적이다. LLM이 실험 조건을 정의하고 원하는 출력을 명확히 하여 복잡한 문제를 단순화하는 방식을 보여준다.
. HuggingGPT: 사용자의 질의를 구조화된 작업 목록으로 변환하고, 실행 순서 및 리소스 종속성을 결정하는 데 LLM을 활용한다.
. CRISPR-GPT: CRISPR 기반 유전자 편집 실험 설계를 자동화하여, 적절한 CRISPR 시스템 선택, 가이드 RNA 설계, 세포 내 전달 방법 추천, 프로토콜 초안 작성, 검증 실험 계획 수립을 지원한다.
. ChemCrow: 반복적 추론(reasoning)과 동적 계획(dynamic planning)을 활용하며, "생각(Thought), 행동(Action), 행동 입력(Action Input), 관찰(Observation)" 루프를 구조적으로 적용하여 실시간 피드백을 기반으로 접근 방식을 정제한다.

. 다중 LLM 시스템(multi-LLM systems) (Coscientist/LLM-RDF): 전문화된 에이전트를 활용하여 문헌에서 실험 방법론을 추출하고, 자연어 설명을 표준화된 프로토콜로 변환하며, 자동화된 플랫폼을 위한 실행 코드를 생성하고, 실행 과정에서 오류를 적응적으로 수정한다.

- **고급 프롬프팅 기법**(상황 내 학습(in-context learning), Chain of Thought, ReAct)이 위 연구들에서 종종 사용되며, LLM 기반 워크플로에서 실험 계획의 신뢰성과 정확성을 향상시키는 데 기여한다. 더욱이, LLM은 반영과 정제(reflection and refinement)를 통해 실험 설계를 지속적으로 평가하고 개선할 수 있는 능력을 갖추고 있다.

. 예) LLM은 전문가 간 토론을 시뮬레이션하여 협력적 대화(collaborative dialogue)를 수행하며, 가설을 검토하고, 반복적 분석을 통해 출력을 개선한다. 이러한 방식은 실제 과학적 문제 해결 과정과 유사하다. (전문가 의견 간의 차이가 문제 공간을 더욱 깊이 탐구하는 계기가 되고, 다양한 관점을 종합하여 엄격한 논의를 거쳐 최적의 합의를 도출하는 과정과 일맥상통)

3.3 실험 과정 자동화(Automating Experimental Processes)

LLM으로 실험 과정에서 반복적이고 시간이 많이 소요되는 작업을 자동화하여, 생산성을 크게 향상시킨다. (데이터 준비, 실험 실행, 분석, 보고서 작성과 같은 노동 집약적인 과정을 LLM 기반 시스템이 수행)

3.3.1 데이터 준비

가장 노동 집약적인 과정인 데이터 정제(cleaning), 라벨링(labeling), 특징 엔지니어링(feature engineering)과 같은 데이터 준비(data preparation) 과정을 LLM이 자동화 할 수 있게 해준다. 또한, 데이터를 확보하기 어려운 상황에서는 LLM이 실험 데이터를 직접 합성(synthesize) 할 수도 있다. 예) 인간 대상 실험을 수행하는 것이 비용이 많이 들거나 윤리적 문제를 초래하는 사회과학(social science) 연구에서 사회적 환경을 시뮬레이션하는 샌드박스(sandbox)를 설계하고, 다수의 LLM 기반 에이전트를 배치하여 서로 상호작용하도록 하였다. 이 방식은 연구자들이 에이전트 간의 사회적 상호작용 데이터를 수집하고 분석할 수 있도록 지원한다.

3.3.2 실험 실행 및 워크플로 자동화

과학 연구에서 실험 워크플로를 자동화하기 위해, LLM 기반 에이전트는 사전 학습(pretraining), 파인 튜닝(fine-tuning), 도구 증강 학습(tool-augmented learning)을 결합하여 작업별(task-specific) 역량을 습득한다. 방대한 데이터 셋을 활용한 사전 학습은 기초 지식을 습득하고, 도메인별 데이터 셋을 활용한 파인 튜닝을 통해 특정 과학적 응용 분야에 맞춰 이 지식을 정교화 한다. 작업 실행(task execution)을 향상시키기 위해, LLM은 종종 도메인별 지식 베이스(domain-specific knowledge bases) 또는 사전 구성된 워크플로(preconfigured workflows)와 결합된다. 또한, 상황 내 학습(in-context learning) 및 Chain-of-Thought 프롬프팅과 같은 고급 프롬프팅 기법을 활용하여, 새로운 실험 프로토콜에 신속하게 적응할 수 있도록 한다. 이와 더불어, 작업별 피드백 루프(task-specific feedback loops)를 통한 반복적 조정에 의해 LLM이 실험 목표에 따라 출력을 지속적으로 정

제할 수 있도록 한다.

이러한 원칙을 바탕으로, LLM은 다음과 같이 다양한 분야에서 실험 워크플로 자동화 역할을 수행한다.

화학

. ChemCrow: LLM 기반 화학 에이전트로, 18가지 전문가 설계 도구를 활용하여 복잡한 화학 합성을 자동으로 계획하고 실행하며, 계산 화학과 실험 화학을 연결한다.
. Coscientist: 실험실 자동화 시스템과 LLM을 통합하여 팔라듐 촉매 합성(palladium-catalyzed synthesis)과 같은 반응을 최적화한다.
. 진화 탐색 알고리즘(evolutionary search strategies): 광범위한 화학 공간을 탐색하여 후보 분자를 식별하고 실험 부담을 줄이는 데 활용된다.
. Bayesian 최적화와 자연어 입력 결합: 촉매 합성(catalyst synthesis)을 위해 자연어 입력과 베이지안 최적화를 결합하여, 반복적 설계 주기(iterative design cycles)를 최적화한다.
. 가설 시나리오 테스트(hypothetical scenario testing) & 반응 설계(reaction design): 사전 스크리닝(pre-screening)을 통해 실험 반복 횟수를 줄인다.

신약 개발

. ChatDrug: 프롬프팅(prompting), 데이터 검색(retrieval), 도메인 피드백(domain feedback) 모듈을 결합하여 신약 설계 및 편집(drug editing)을 지원한다.
. DrugAssist: 인간-기계 대화(human-machine dialogue)를 통해 분자 구조를 반복적으로 최적화(iterative optimization) 한다.

생물학 및 의학

. ESM-1b & ESM-2: 단백질 서열을 인코딩하여 이차 및 삼차 구조 예측과 같은 예측 작업을 수행한다.
. 단백질 생성 모델(protein generation model): 단백질 패밀리 데이터로 LLM을 파인 튜닝(fine-tuning)하여, 기능적으로 유효하면서도 다양한 단백질 서열을 생성한다.
. 항체 생성 LLM(antibody generative LLM): 신규 SARS-CoV-2 항체 디자인을 위해 개발되었다. 자연 항체에 의존하지 않고도 특이성(specificity)과 다양성(diversity)을 갖춘 항체를 설계한다.

3.3.3 데이터 분석과 해석

LLM은 데이터 분석에 대한 자연어 기반 설명(natural language explanations)을 생성하고, 의미 있는 시각화(visualizations)를 구성하여 복잡한 데이터 셋의 해석에 기여한다. 도출된 인사이트(insights)를 보다 이해하기 쉽고, 실행 가능하도록(accessible and actionable) 만드는 역할을 수행한다. 전통적으로 데이터 분석(data analysis)은 광범위한 통계적 전문 지식, 수작업 계산, 방대한 실험 결과 해석을 필요로 하는데, LLM은 통계 모델링(statistical modeling) 및 가설 검정과 같은 작업을 자동화하여 분석 과정을 단순화 시켜준다.
예) LLM이 모델러(modeler)로서, 확률 모델(probabilistic models)을 제안하고, 적합(fitting) 및 정제하며, 또한 후방 예측 점검(posterior predictive checks)과 같은 기법을 활용하여 모

델 성능에 대한 피드백을 제공해준다.

- 패턴, 트렌드, 관계 탐색: LLM은 텍스트 데이터에서 숨겨진 패턴(hidden patterns), 트렌드(trends), 관계(relationships)를 발견하는 데 뛰어난 성능을 발휘.
- 소셜 미디어 데이터 분석: 공공 여론(public sentiment)과 신흥 트렌드(emerging trends) 파악.
- 환경 데이터 해석: 환경 과학(Environmental Science) 분야에서 의사결정을 지원.
- 주제 분석(Thematic Analysis): 질적 데이터(qualitative data)에서 주요 주제와 패턴을 식별.
- 금융 예측(Financial Forecasting) 및 위험 평가(Risk Assessment): 데이터 기반 금융 분석을 강화.
- 다중 에이전트 기반 데이터 분석 프레임워크 AutoGen: 여러 개의 사용자 지정 가능한 에이전트(customizable agents)를 활용하여, 다양한 LLM 기반 애플리케이션을 구축할 수 있는 범용 프레임워크(generic framework)를 제공. 이 에이전트(customizable agents)들은 자연어 및 코드로 상호 작용하며, 데이터 모델링(data modeling) 및 데이터 분석(data analysis)과 같은 다양한 다운스트림 작업을 지원.

3.5 도전 및 향후 연구

도전 과제

모델 자체의 내재적 한계와 특정 도메인에 적용할 때 발생하는 문제에서 비롯된다.

① 계획 능력의 한계(Planning Capability Limitations)

LLM의 계획 수립 능력(planning capability)에는 근본적인 한계가 존재한다. LLM은 자율 실행 모드에서 실행 가능한 계획(executable plans)을 생성하는 데 실패하는 경우가 많은데, 이는 할루시네이션(hallucination, 환각 현상)에 취약하기 때문이다. 즉, 비합리적인 계획을 생성하거나, 주어진 작업 지침(task prompts)에서 벗어나는 오류를 범하거나, 복잡한 지침을 따르는 데 어려움을 겪을 가능성이 높다.

② 프롬프트 강건성(Prompt Robustness)의 문제

다단계 실험(multi-stage experiments)에서 프롬프트 강건성(prompt robustness)이 중요한 도전 과제이다. 같은 의미를 전달하더라도 프롬프트 문구의 사소한 차이가 실험 계획 및 실행 과정 전반에 걸쳐 일관성을 해칠 수 있다. 이러한 문제는 실험 결과(experimental outcomes)에 직접적인 영향을 미칠 수 있다.

③ 처리 속도 문제(Slow Processing Speed)

자기회귀적 LLM(autoregressive LLMs)의 느린 처리 속도는 실시간 피드백(real-time feedback)을 필요로 하는 반복적(iterative) 및 다단계(multi-step) 실험 계획에서 병목현상을 유발할 수 있다.

④ 특정 도메인 적용의 어려움(Application-Specific Challenges)

LLM은 특정 연구 분야의 전문 지식 및 인지 과정(cognitive processes)을 완벽하게 모방하는 데 어려움을 겪기 때문에, 다양한 연구 분야에서 일반화할 수 있는 능력에 한계가 존재한다. 예) 윤리적으로 민감하거나(ethically sensitive), 오류 발생 가능성이 높은 실험을 시뮬레이션할 필요가 있는 경우, LLM이 내재적으로 포함하고 있는 안전성 준수 원칙(safety-aligned values)과 충돌할 가능성이 높다.

향후 연구 방향

핵심 모델 역량을 강화하고, 실험 작업의 고유한 요구사항에 맞게 모델을 최적화하는 것이 중요하다. 할루시네이션(hallucination) 문제를 완화하기 위해, 워크플로우에 강력한 검증 메커니즘(robust verification mechanisms)을 통합하는 방안을 고려할 수 있다. 예) 외부 신뢰 가능한 검증 시스템(external sound verifiers)과 결과를 교차 검증하거나, 실시간 피드백 루프를 활용하여 오류를 동적으로 수정하는 방법이 있다.

- 프롬프트 강건성(prompt robustness) 개선을 위해, 맥락적 변화(contextual changes)에 따라 프롬프트 구조를 모니터링 및 조정하는 적응형 시스템 개발이 필요하다. 이를 통해 실험 계획 단계 전반에서 일관성을 유지할 수 있다.
- 모델의 효율성(efficiency) 향상을 위해, 다단계 추론(multi-step reasoning)에 최적화된 증류(distilled) LLM을 개발하거나, LLM과 업무특화(task-specific) 모델을 결합한 하이브리드 시스템(hybrid systems)을 구축하여 속도와 정확성 간 균형을 유지하는 방안을 고려할 수 있다.
- 특정 역할 적응(role adaptation) 능력을 강화하기 위해, 고품질 도메인 특화 데이터 셋으로 LLM을 정밀 튜닝하거나, 모듈형 프레임워크를 개발하여 특정 과학적 추론을 더욱 정밀하게 모방할 수 있도록 해야 한다.
- 윤리적으로 복잡한 시나리오에서 LLM이 안전하게 작동할 수 있도록, 적응형 정렬 프로토콜(adaptive alignment protocols)을 설계하여 특정 실험 목표에 맞게 LLM을 조정하는 연구가 필요하다.

4. LLM에 의한 과학 논문 작성

4.1 개요

LLM에 의한 인용문 생성, 관련 연구 정리, 초안 및 본문 작성에 사용된 방법론, 모델의 효과성, 그리고 과학적 글쓰기 자동화에서 직면하는 도전 과제를 분석한다.

4.2 인용문 생성

인용문 생성(citation text generation) 작업을 통해 참조할 논문들의 정확한 텍스트 요약을 생성한다. LLM은 풍부한 문맥 이해력과 일관성을 바탕으로 인용문 생성을 자동화하는 데 중요한 역할을 하며, 정확성과 활용도를 높이기 위한 다양한 방법론을 적용하고 있다.

. Xing et al.: 포인터-생성 네트워크(pointer-generator network)를 활용, 교차 주의 메커니즘(cross-attention mechanisms)을 통해 인용 논문의 원고 및 초록에서 단어를 복사하여 인용문을 생성한다.

. Li와 Ouyang: LLM을 프롬프트하여, 인용 네트워크 내 논문 쌍(pairs of papers) 간의 관계를 강조하는 자연어 설명을 생성하도록 한다.

. AutoCite 및 BACO: 이 연구를 확장하여 다중 모드(multimodal) 접근법을 적용, 인용 네트워크 구조와 텍스트 문맥을 결합하여 문맥적으로 적절하고 의미적으로 풍부한 인용문을 생성한다.

. Gu와 Hahnloser, Jung et al.: 사용자가 인용 의도 및 키워드와 같은 속성을 지정할 수 있도록 허용하고, 이를 구조화된 템플릿에 통합한 후, 언어 모델(LM)을 미세 조정하여 사용자의 요구에 맞는 인용문을 생성한다.

4.3 관련 연구 정리

LLM을 활용하여 최신 참조 논문들을 바탕으로 과학 논문의 관련 연구 내용을 작성한다. LLM이 다양한 자연어 이해 및 생성 작업에서 성공을 거두었고, 최근에는 보다 포괄적이고 세밀한 문헌 리뷰가 가능해졌으며, 이는 다양한 연구 분야 간 더 깊은 통찰력과 연결을 촉진하고 있다.

- Martin-Boyle et al., Zimmermann et al.: ChatGPT를 활용하여 문헌 리뷰 작업과 관련 연구 생성에 대한 사례 연구(case studies)를 하였다. ChatGPT가 과학 논문 데이터셋을 빠르게 스캔하고 관련 연구 섹션의 초기 초안(initial draft)을 생성하는 데 도움이 됨을 보여주었다.
- 할루시네이션(hallucinations) 문제: 생성된 콘텐츠가 사실에 기반 하지 않거나 최신 연구를 정확하게 반영하지 못하는 문제가 발생할 수 있다. 이러한 문제를 해결하기 위해, 많은 연구들은 검색 증강 생성(Retrieval-Augmented Generation, RAG) 원칙을 적용하였다, RAG는 외부 출처에서 검색된 사실 기반 콘텐츠를 바탕으로 LLM 기반 문헌 리뷰 생성을 강화하는 방법이다.

. LitLLM: RAG를 활용하여 웹사이트에서 관련 논문을 검색하고 재정렬하여, 포괄적인 문헌 리뷰에 필요한 시간과 노력을 줄이는 동시에 할루시네이션을 최소화하였다.

. HiReview: 이를 더욱 발전시켜, RAG 기반 LLM을 그래프 기반 계층적 클러스터링과 통합하였다. 이 시스템은 먼저 인용 네트워크 내에서 관련 하위 커뮤니티를 검색하고 계층적 분류 트리(hierarchical taxonomy tree)를 생성한 후, LLM은 각 클러스터에 대해 요약을 생성, 완전한 커버리지와 논리적 구성을 보장하였다.

- Nishimura et al.: 관련 연구 섹션에서 참신성(novelty)을 강조하도록 LLM을 통합하였다. 새로운 연구와 기존 연구를 비교하여, LLM은 새롭고 다른 점을 명확하게 강조하는 관련 연구 섹션을 생성하는 데 도움을 주며, 대상 논문과 이전 문헌 간의 비교를 해준다.

4.4 초안 및 본문 작성

자동화된 과학적 글쓰기 분야에서 LLM은 특정 텍스트 요소 생성부터 전체 연구 논문 작성에 이르는 다양한 작업에 사용되고 있다.

. August et al.: 다양한 사용자에 맞춘 제어 가능한 복잡도의 과학적 정의 생성을 제안한다.

. SCICAP: 과학적 그림의 캡션 자동 생성으로 시각적 데이터를 빠르고 정확하게 설명한다.

. PaperRobot: 더 포괄적인 시스템으로써, 점진적인 초안 작성법을 도입하여 사용자 입력을 바탕으로 논문의 각 섹션을 조직하고 초안을 작성하는 데 LLM을 활용한다.

. CoAuthor: 협업적 인간-AI 접근법을 취하여, LLM이 저자에게 제안 생성 및 텍스트 확장을 통해 도움을 준다.

. Ifargan et al.: 완전 자율적인 글쓰기를 위해 LLM이 데이터 분석부터 최종 초안까지 전체 연구 논문을 생성하는 방법을 탐구하였다.

. AutoSurvey: LLM이 기존 연구를 종합하고 자율적으로 종합적인 설문을 작성한다.

. AI Scientist와 CycleResearcher: 과학적 논문 초안 작성뿐만 아니라, 가설 생성 및 실험 설계를 포함한 전체 과학적 과정에 기여하는 시스템을 제안하여, 완전 자동화된 과학적 발견 및 글쓰기의 가능성을 강조한다.

4.6 도전 및 향후 과제

도전 과제(Challenges)

사실의 정확성 유지, 맥락의 일관성 보장, 복잡한 정보 처리 어려움과 같은 LLM의 내재적 한계로 인해 인용문 생성, 관련 연구 작성, 초안 및 논문 작성에 도전과제가 나타난다.

- 환각(hallucinations)으로 인해 잘못된 인용이나 관련 없는 인용을 생성하는 경우가 많고, 또한, 의존하는 검색 시스템에 의해 제약을 받는다.
- 제한된 문맥 창(context window)은 모델이 방대한 참조를 관리하거나 관련 문헌을 포괄적으로 통합하는 능력을 더욱 제한한다. 이로 인해 잘못된 인용 순서나 부적절한 인용 그룹화가 발생할 수 있다. 또한, 과학적 엄밀성을 보장하고 피상적이거나 사소한 출처에 의존하지 않도록 하는 것은 지속적인 장애물로 남아 있으며, LLM은 학술적 글쓰기에 필요한 깊이와 논리를 제공하는 데 어려움을 겪고 있다.
- LLM을 학술적 글쓰기에 사용하는 것은 학문적 무결성 및 표절과 관련된 중요한 윤리적 문제를 제기한다. 이는 저자 표시(authorship)의 경계를 모호하게 한다. 또한, 연구자들이 기계 생성 텍스트를 자신들의 작업에 의한 문장으로 발표할 수 있다. LLM은 기존 문헌을 유사하게 모방하는 텍스트를 생성할 수 있기 때문에, 생성된 콘텐츠가 충분히 독창적이지 않아 의도하지 않은 표절의 위험을 초래할 수 있다.
- 논문 초안을 작성하는 데 LLM을 사용하는 편리함은 전통적으로 학술 글쓰기에서 요구되는 지적 노력을 저하시킬 수 있으며, 이로 인해 학문적 연구에 필수적인 학습 과정과 비판적 사고 능력이 저평가될 위험이 있다.

향후 연구 방향(Future Work)

도전 과제 극복을 위해, 검색 시스템 개선과 다양하고 긴 문맥 출처에서 정보를 합성하는 모델의 능력 향상에 초점을 맞추어야 한다. 여기에는 더 나은 인용 검증 메커니즘 개발, 다중 문서 합성 향상, 그리고 실시간 문헌 발견을 도입하여 생성된 콘텐츠를 최신 상태로 유지하는 방법이 포함된다.

- 도메인별 미세 조정과 추론 인식 모델을 통합하면 보다 정확하고 맥락에 맞는 과학적 텍스트 생성을 돕는다. 글쓰기 과정에 대한 세밀한 제어(예: 톤과 스타일 조정)는 LLM이 다양한 학술적 요구에 적응할 수 있도록 개선하는 데 중요하다. 더 나아가, 인간이 참여하는 시스템(Human-in-the-loop 시스템)-인간의 감독과 개입이 글쓰기 과정에 필수적인 분야에-을 통합하면 학문적 작업에 내재된 지적 엄밀성 및 비판적 사고를 보장할 수 있다.
- 윤리적 문제를 해결하기 위해 학계는 LLM 사용에 대한 명확한 지침과 윤리적 기준을 설정하여 학술 작업의 무결성과 독창성을 보장하는 것이 중요하다.

5. 동료 심사를 위한 LLM

5.1. 개요

LLM과 peer review 과정 통합은 리뷰어 편향, 일관되지 않은 기준, 업무 부담 불균형과 같은 오랜 문제를 해결하는 엄청난 진전이 있음을 나타낸다.

- LLM을 활용한 검토 방식 도입 예: ICLR 2025는 리뷰어들의 심사 과정을 지원하기 위해 LLM 기반 시스템을 도입할 것이라고 발표했다.

두 가지 주요 접근 방식으로 발전

자동 리뷰 생성: 논문 제출 증가에 따른 리뷰어의 업무 부담을 줄이려는 필요성에서 등장했

다. LLM이 연구 논문을 독립적으로 분석하여, 방법론 검증, 결과 확인, 기여도 평가 등의 여러 측면을 평가하고, 인간의 직접적인 개입 없이 포괄적인 리뷰 보고서를 제공하도록 설계되었다.

LLM 보조 리뷰 워크플로우: 학문적 평가에서 인간 전문가의 역할이 여전히 중요하다는 인식을 바탕으로 LLM이 보조 도구로 활용되며, 논문 요약, 참고문헌 검토, 내부 일관성 검사와 같은 시간이 많이 걸리지만 명확하게 정의된 작업을 지원한다. 반면, 핵심적인 평가 및 판단은 여전히 인간 심사자가 담당한다.

이 접근 방식들은 리뷰의 효율성, 일관성, 품질을 향상하기 위해 다양한 방법론을 적용학고 있다. 또한, 리뷰 시스템을 체계적으로 평가하고 개선하기 위해, 표준화된 학습 데이터 셋을 제공하고 성능 평가 지표를 설정하는 특수한 peer review 벤치마크가 개발되었다. 본 장에서는 이러한 방법론과 평가 프레임워크를 분석하고, 구현상의 도전 과제 및 향후 연구 방향을 살펴본다.

5.2. 자동 리뷰 생성(Automated Peer Review Generation)

최소한의 인간 개입으로 LLM이 포괄적인 리뷰를 생성할 수 있도록 하여 심사 과정을 간소화하는 것을 목표로 한다. 이 시스템들은 연구 논문을 입력받아 전체적인 peer review 또는 메타 리뷰(meta-review)를 생성하며, 피드백의 깊이, 정확성, 관련성을 향상시키기 위한 다양한 기법을 활용한다.

A. 단일 모델 접근법

정교한 프롬프트 기법과 모듈식 설계를 활용하여 review 생성 프로세스를 최적화하는 데 초점을 맞춘다. 이 방식에서는 신중하게 설계된 프롬프트를 사용하여 모델이 논문의 특정 요소(예: 방법론, 연구 결과, 기여도 등)에 집중하도록 유도한다.

이 접근법 내에서도 다양한 아키텍처가 제안되었다.

. MetaGen: 요약 추출 과정 후 의사결정 중심의 세부 보완 과정의 2단계 파이프라인 사용

. Kumar et al.: 의사 결정 예측과 검토 생성을 통합하는 신경망 아키텍처 개발

. MReD: 문장 수준의 기능 라벨을 활용한 구조화된 생성 기법 도입

. CGI2: 위 연구를 기반으로 모듈식 설계를 통해 단계별 리뷰 프로세스를 구현한다.

  * 논문의 주요 의견을 추출
  * 연구의 강점과 약점을 요약
  * 체크리스트 기반 프레임워크를 활용하여 반복적 피드백을 통해 검토 품질을 정제

  . 이러한 반복적 접근 방식은 리뷰의 깊이와 관련성을 향상시키지만, 지나치게 복잡한 방법론을 다루는 논문이나 컨텍스트 윈도우를 초과하는 긴 논문에서는 한계를 보인다.

. CycleReviewer: 강화 학습을 활용하여 검토 품질을 지속적으로 개선하는 엔드투엔드 리뷰 생성 방식을 채택. (강점) 리뷰의 정확성과 명확성을 향상. (한계점) 높은 계산 비용이 요구되어 확장성 문제 발생

. ReviewRobot: 지식 그래프(knowledge graph)를 활용하여 논문의 핵심 내용을 구조적으로 분석하고, 이를 기반으로 상세한 리뷰 코멘트를 생성. (강점) 뛰어난 설명 가능성과 증거 기반 평가 제공. (한계점) 미리 정의된 템플릿을 기반으로 하기 때문에 유연성이 부족

B. 다중 모델 접근법

여러 개의 특화된 모델을 조합하여 검토 과정의 각 단계를 처리하는 방식이다. 이는 복잡한 논문을 보다 효과적으로 다룰 수 있으며, 특정 전문 분야에 대한 심층적인 평가가 가능하다.

. Reviewer2: 두 단계로 진행되는 리뷰 프로세스 구현. 한 모델이 특정 리뷰 포인트(프롬프트)를 생성하고 다른 모델이 이를 활용해 상세하고 집중적인 피드백 작성. (강점) 보다 세밀하고 정교한 피드백 가능. (한계점) 통합적 프레임워크 부족으로 인해 부분적이거나 편향된 리뷰가 발생할 가능성 존재.

. SEA: 검토 과정에서 표준화, 평가, 분석을 담당하는 개별 모델을 활용하여 보다 균형 잡힌 리뷰를 제공. (핵심 기능) 여러 개의 리뷰를 통합하여 일관성을 높이고 중복을 줄임. 논문과 생성된 리뷰 간의 불일치를 측정하는 미스매치 점수(mismatch score) 도입. 자기 수정(self-correction) 전략을 활용하여 반복적으로 리뷰 품질 개선. (강점) Reviewer2보다 일관성과 포괄성이 뛰어남. (한계점) 여러 모델 간의 조정이 필요하여 시스템 복잡성이 증가

. MARG: 일반적인 LLM 컨텍스트 제한을 초과하는 긴 논문을 처리하기 위한 멀티 에이전트 프레임워크 도입. 여러 개의 특화된 모델이 역할을 나누어 각 섹션을 개별적으로 검토. 긴 논문에서도 세부적인 피드백을 유지할 수 있도록 설계. (강점) 특정 영역별 세밀한 피드백 제공 가능. (한계점) 여러 에이전트 간 출력 조정 및 정합성 유지 문제가 발생할 가능성.

단일 모델: 구현이 간단하고 제어가 용이하나, 긴 논문이나 복잡한 연구에 한계 지님.
다중 모델: 확장성이 뛰어나고 특정 검토 작업에 최적화 가능하나 요소들 간의 세밀한 조정이 요구되고 일관성 유지가 어려움.

예)

. ReviewRobot: 구조화된 접근 방식을 통해 설명 가능성과 실행 가능한 인사이트를 제공하지만, 변화하는 연구 환경에 적응하는 능력이 떨어진다.

. CycleReviewer: 상당한 계산 자원 없이 반복정제를 통해 동적 적응력을 향상시킨다.

이러한 한계를 고려할 때, 단일 모델의 단순성과 다중 모델의 적응성을 결합하는 하이브리드 접근 방식이 리뷰 품질, 일관성, 포괄성을 모두 향상시키는 유망한 방향이 될 수 있다.

5.3 LLM 보조 리뷰 워크플로우(LLM-assisted Peer Review Workflows)

LLM을 활용한 피어 리뷰 워크플로우는 인간 심사자의 역량을 향상시키는 데 초점을 맞춘다. 최근 연구들은 인간-AI 협업 접근 방식이 매우 중요함을 나타내고 있다. LLM이 효율성을 향상시킬 수는 있지만, 윤리적 기준을 유지하고 리뷰의 신뢰성을 보장하기 위해서는 여전히 인간의 감독이 필수적이다.

예) AgentReview: LLM이 초기 인사이트를 생성하면 인간 리뷰어가 이를 정제하고 검증하는 방식으로 운영된다.

LLM 보조 피어 리뷰 워크플로우의 세 가지 주요 기능

(1) 정보 추출 및 요약: 리뷰어가 논문의 내용을 빠르게 파악할 수 있도록 돕는다.
(2) 원고 검증 및 품질 보증: 논문의 주장을 체계적으로 검토하고 검증하는 데 도움을 준다.
(3) 리뷰 작성 지원: 잘 구성된 피드백을 생성하는 과정을 보조한다.

(1) 정보 추출 및 요약

시스템이 문서 이해 및 요약을 자동화하여 리뷰어의 논문 이해를 돕는다.

. PaperMage: 자연어 처리(NLP) 및 컴퓨터 비전 모델을 통합한 기본 툴킷으로, 논문의 논리적 구조, 도표, 텍스트 콘텐츠를 정밀하게 분석하여 시각적으로 복잡한 과학 문서를 처리한다.

. CocoSciSum: 논문의 핵심 내용을 요약하는 데 초점을 맞추며, 길이 조절 및 특정 키워드 포함 여부를 사용자가 지정할 수 있도록 지원한다. 또한, 해당 모델은 정밀한 정보 조합 방식(compositional control architecture)을 활용해 높은 사실 정확도를 유지한다.

(2) 원고 검증 및 품질 보증

시스템이 다양한 수준에서 과학적 엄밀성을 검토한다.

- **로컬 수준(Local Level):** ReviewerGPT는 체계적인 오류 탐지 및 가이드라인 준수를 전문적으로 수행하며, 논문 제출 요건 검증뿐만 아니라 개별 원고 내의 수학적 오류 및 개념적 불일치를 효과적으로 식별할 수 있다.
- **글로벌 수준(Global Level):** PaperQA2는 논문의 주장과 기존 과학 문헌을 비교하여 검증하는 방식으로 작동하며, 고급 언어 에이전트를 활용해 모순을 감지하고 진위를 판별한다. 해당 시스템은 논문 당 평균 2.34건의 검증된 모순을 식별하며, 문헌 간 분석에서도 높은 사실 정확도를 유지한다.
- **아이디어 검증(Idea Validation):** Scideator는 측면 결합(facet recombination)을 통해 논문의 개념을 검토하고 과학적 유사성을 찾아내는 방식으로 작동한다. 또한, 참신성 검사(novelty checker)를 포함하여, 연구 주장이 기존 연구 패러다임에 적절히 부합하는지 평가할 수 있도록 지원한다.

(3) 리뷰 작성 지원

시스템이 리뷰어의 전문성 수준에 따라 다양한 방식으로 보조한다.

- **개별 리뷰어 지원:** ReviewFlow는 컨텍스트 반영을 위한 힌트와 노트 정리 가이드를 제공하여, 초보 리뷰어가 논리적으로 구성된 리뷰를 작성할 수 있도록 돕는다. 이 시스템은 피어 리뷰 과정을 단계별로 분해하여, 복잡한 작업을 보다 쉽게 수행할 수 있도록 지원한다.
- **협업 리뷰 작성 지원:** CARE는 자연어 처리(NLP)를 활용한 인라인 주석 및 실시간 협업 기능을 제공하는 통합 플랫폼으로, 리뷰어들이 보다 효과적으로 협력하면서 상세하고 건설적인 피드백을 제공할 수 있도록 돕는다.
- **문서 워크플로우 자동화:** DocPilot은 모듈식 작업 계획 및 코드 생성 기능을 활용하여 문서 처리 과정의 반복적이고 복잡한 작업을 자동화한다. 이를 통해 과학 논문의 PDF를 체계적으로 관리하고 주석을 달 수 있도록 지원하여, 리뷰어가 절차적 문제보다 실질적인 피드백에 집중할 수 있도록 한다.

5.5 도전 및 향후 과제

도전 과제

동료 심사의 핵심은 깊이 있는 전문성, 미묘한 차이 이해, 그리고 신중한 판단에 있다. LLM이 보이는 한계는 학문적 평가의 자동화가 얼마나 복잡한지를 드러낸다.

기술적 한계 문제

- LLM이 특정 학문 분야의 전문 용어나 복잡한 개념을 완전히 이해하는 데 어려움을 겪는다. 이러한 제한된 기술적 이해는 연구 방법론 평가 능력에도 직접적인 영향을 미친다. 특정 분야의 개념을 충분히 이해하지 못하면, 연구 방법이 적절한지, 증거가 결론을 뒷받침하는지

신뢰할 수 있는 평가를 내리기 어렵다. 예) 학제 간 연구(interdisciplinary research)에서는 분야별 방법론적 기준이 다를 수 있는데, LLM은 표본 크기의 부적절성, 잘못된 통계 검정, 실험 통제의 결여와 같은 중요한 문제를 식별하지 못하는 경우가 많다. 이는 연구의 질과 과학적 무결성을 보장하는 동료 심사의 중요한 역할을 고려할 때 심각한 문제가 된다.

- 학술적 글쓰기의 복잡성도, LLM이 긴 논문을 다룰 때, 추가적인 난관을 제기한다. LLM은 문맥 창(context window)이 확장되더라도 일관된 분석을 유지하는 데 어려움을 겪으며, 여러 섹션에 걸친 복잡한 논증을 놓치는 경우가 많다. 더욱 우려되는 것은 "환각(hallucination)" 문제로, 모델이 설득력 있지만 잘못된 평가를 생성하는 경우가 있으며, 특히 새로운 연구 방법론을 검토할 때 이러한 문제가 자주 발생한다.

기술적 한계를 넘어서는 추가적인 문제

- 전문적인 훈련 데이터 부족: 한문 분야 전반에 걸친 불균형 환경(uneven landscape)을 조성하며, 연구 공동체 규모가 작거나 특수한 용어를 사용하는 분야에서 더욱 두드러진다.
- LLM 활용과 관련한 윤리적 문제: 알고리즘적 편향과 투명성 문제 외에도, "표절 세탁(plagiarism laundering)"과 같은 새로운 형태의 학문적 부정행위가 등장하고 있다. 또한, 다수의 연구자가 동일한 LLM 시스템을 사용해 심사를 진행할 경우, 학문적 피드백이 획일화될 위험도 있다. 동료 심사의 핵심 요소 중 하나는 개별적이고 다양한 관점에서 비롯되는 창의적 통찰인데, AI 도구의 광범위한 사용이 이러한 다양한 창의적 통찰을 감소시키고 다양한 관점을 저해할 가능성이 있다.

향후 연구 방향

- 학술 논문 심사에서 LLM의 역량을 향상시키기 위한 근본적인 기술적 문제를 해결하는 것이 우선적으로 되어야 한다. 이를 다음과 같이 세가지로 정리할 수 있다.
- 첫째, LLM은 다양한 학문 분야에서 사용되는 전문적인 기술 개념을 이해하는 데 어려움을 겪으므로, 특정 분야의 용어를 보다 효과적으로 처리하고 이해할 수 있는 방법이 필요하다.
- 둘째, 인용 분석(citation analysis) 기능을 강화하여 참고문헌의 관련성을 검증하고, 논문의 논증을 얼마나 효과적으로 뒷받침하는지 평가할 수 있어야 한다.
- 셋째, 긴 학술 문서를 분석하는 과정에서 일관성을 유지할 수 있도록, 섹션 간 교차 참조를 통한 연계 분석 및 연구 방법·결과·결론 간 일관성 검증을 위한 새로운 방법론이 요구된다.

기술적 개선 외의 과제

- 효과적인 인간-AI 협업 프레임워크를 개발하는 것이 중요하다. 차세대 심사 시스템은 연구자가 잠재적 문제를 식별하고, 기존의 연구 심사 워크플로우와 원활하게 통합될 수 있는 직관적 인터페이스를 제공해야 한다. 또한, 이러한 협업 시스템은 다양한 학문 분야에서 활용될 수 있도록 설계되어야 한다. 인간-AI 시스템의 효과성을 검증하기 위해서는 엄격한 평가 프레임워크를 구축하여, 실제로 연구자의 심사 효율성과 효과성을 향상시키는지를 평가해야 한다.
- LLM의 동료 심사 도입이 점점 보편화됨에 따라, 신뢰할 수 있는 거버넌스 체계를 마련하는 것이 필수적이다. 여기에는 LLM이 생성한 콘텐츠를 감지하고, LLM의 기여를 투명하게 추적하며, 심사자의 신뢰성을 유지하는 신뢰성 높은 방안이 포함된다. 또한, 기존 저널 플랫폼과 LLM 심사 도구를 안전하게 통합하기 위한 표준화된 프로토콜이 필요하다.

종합적인 평가 프레임워크 마련

- 기술적 역량 평가: 언어 이해, 인용 분석, 문서 일관성 측면에서의 향상을 체계적으로 측정

해야 한다.

- 인간-AI 협업 지표: LLM이 제공하는 피드백의 품질과 심사자 효율성에 미치는 영향을 평가해야 한다.
- 거버넌스 측면: LLM 감지 시스템의 신뢰성과 플랫폼 통합의 보안성(security)을 검증해야 한다. 특히, 다양한 학문 분야, 출판 형식, 언어적 배경에 따른 편향성을 분석하여, 모든 학술 공동체에 공정한 지원이 이루어질 수 있도록 해야 한다.
- 이러한 평가를 기반으로, 동료 심사 과정의 신뢰성을 유지하면서도 LLM 시스템이 실질적으로 기여할 수 있는 방향으로 발전할 수 있도록 해야 한다.

6. 결론

결론적으로, LLM은 현대 과학 연구의 모든 단계에서 새로운 방법을 제공하는 첨단의 생산적 도구로 자리 잡고 있다. 특정 분야의 과업에서는 내재된 한계, 기술적 장벽, 윤리적 고려 사항이 존재하지만, LLM 기술이 지속적으로 발전하면서 연구 방식에 혁신을 가져올 것으로 기대된다. 이러한 시스템이 발전함에 따라 과학적 워크플로우에 통합되어 연구 속도를 가속화할 뿐만 아니라, 과학 공동체 내에서 전례 없는 혁신과 협업을 촉진할 것이다.

LLM 활용 과학 연구 조사:AI Co-Scientist

취지: 구글의 Co-Scientist(부과학자)는 LLM 활용 과학 연구의 대표적인 실적으로 평가받고 있다. 이에 구글의 Co-Scientist가 어떻게 활용되고 있으며, 어떠한 한계가 존재하고, 어떠한 방향으로 나아가고 있는지를 살펴보고자 한다.

“Towards an AI co-scientits”

Juraj Gittweis et al., https://arxiv.org/abs/2502.18864

과학적 발견은 과학자들이 새로운 가설을 생성하고 이를 엄격한 실험적 검증을 거치는 과정에 의해 이루어진다. 이 과정을 보완하기 위해 개발된 Gemini 2.0을 기반으로 한 다중 에이전트 시스템 “AI Co-Scientist”를 살펴본다. AI Co-Scientist는 기존의 증거를 바탕으로 연구자의 연구 목표와 지침에 맞춰 새로운 원천 지식을 발견하는 데 도움을 주고 새로운 연구 가설과 제안을 생성하도록 설계되었다.

AI Co-Scientist는 가설 생성 과정에서 생성(generation), 토론(debate), 발전(evolution)이라는 접근 방식을 적용한 과학적 방법론을 반영하며, 테스트 시점에서의 연산 확장을 통해 가속화된다.

주요 기여 사항은 다음과 같다.

(1) 유연한 연산 확장을 지원하는 비동기적 작업 실행 프레임워크를 갖춘 다중 에이전트 아키텍처

(2) 스스로 개선되는 가설 생성을 위한 토너먼트 진화 과정

자동화된 평가 결과, 테스트 시점에서의 연산 확장 효과에 의해 지속적 가설 품질 향상이 이루어짐을 확인하였다. 이 시스템은 범용적으로 활용될 수 있지만, 우선적으로 생물의학 분야의 세 가지 영역—① 약물 재창출(drug repurposing), ② 신규 표적 발견(novel target discovery), ③ 박테리아 진화 및 항미생물 저항성의 메커니즘 설명—에서 개발 및 검증을 진행하였다.

1 서론

- 인간의 창의성과 독창성은 과학의 기초 연구 발전을 이끌어왔지만, 현재의 연구자들은 폭과 깊이의 딜레마에 직면해 있다. 연구의 복잡성은 점점 더 깊고 세부적인 전문 지식을 요구하는 반면, 획기적인 통찰은 여전히 여러 학문을 가로지르는 광범위한 지식의 융합에서 나온다. 그렇지만, 과학 논문의 급격한 증가와 다양한 고속 분석 기술에 의해 특정 분야의 깊이 있는 전문성과 학제 간 통찰력을 동시에 습득하는 것이 점점 더 어려워지고 있다.
- 현대의 많은 혁신적인 돌파구들은 학제 간 협력을 통해 이루어졌다.

. 2020년 노벨 화학상 “CRISPR 연구”: 미생물학, 유전학, 분자생물학 등 여러 분야의 결합

. 2024년 노벨 물리학상 “Artificial Neural Networks”: 물리학과 신경과학의 개념을 결합

- 인공지능(AI)은 General Intelligence를 갖춘 협력적인 시스템으로 빠르게 발전하면서 과학자들이 학문 간 경계를 창의적으로 넘나들고 전문적으로 사고할 수 있도록 하고 있다. 이러한 AI 시스템은 고도화된 추론 능력, 멀티모달 이해력, 그리고 복잡한 문제를 장기적으로 해결하기 위한 도구 활용과 같은 자율적(agentic) 행동을 수행한다. 또한, 지식 증류(distillation) 및 추론 시점 연산 비용(inference-time compute cost) 절감으로 지능적이고 범용적인 AI 시스템이 점점 더 저렴하게 접근 가능해지고 있다.
- 현대 과학 및 의학 연구의 미해결 사항을 위해 AI Co-Scientist 시스템을 개발하였다.

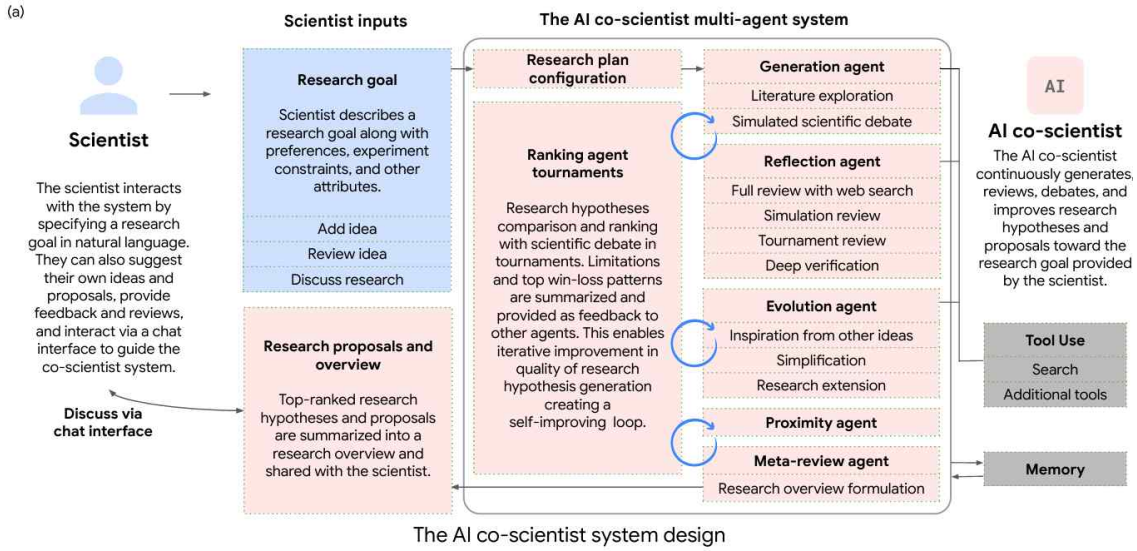


그림 19 . AI Co-Scientist 시스템.

AI Co-Scientist (연구자들의 유용한 조력자 및 협력자)

- Gemini 2.0을 기반으로 한 다중 에이전트 AI 시스템으로, “과학자 중심 협력 모델(scientist-in-the-loop collaborative paradigm)”을 기반으로 과학적 방법론을 반영하도록 설계되었다. 연구자가 간단한 자연어로 연구 목표를 지정하면, 이 시스템은 관련 문헌을 검색하고 논리적으로 분석하여 기존 연구를 요약 및 종합한 후, 이를 바탕으로 새롭고 독창적인 연구 가설과 실험 프로토콜을 제안한다(Figure 1a). 또한, 제안의 타당성을 보장하기 위해 관련 문헌을 인용하고, 그 근거를 설명한다.
- 연구자들은 AI가 생성하는 가설이나 연구 제안에 원하는 특성과 제약 조건을 지정할 수도 있다. 연구자의 아이디어와 가설을 제공하고, AI가 생성한 가설을 직접 수정하거나, 자연어 채팅을 통해 AI의 방향성을 조정하는 등 다양한 방식으로 협업할 수 있다. 테스트 시점 연

산 확장(scaling of test-time compute)을 대폭 활용하여, 점진적으로 논리적 사고를 수행하고, 가설을 발전시키며, 점점 더 정교한 결과를 도출한다.

- 시스템의 **핵심 알고리즘**은 다음과 같은 과정으로 이루어진다.

*** 자기 대결(self-play) 기반 과학적 토론 과정을 통해 새로운 연구 가설을 생성**

*** 토너먼트 방식 평가를 통해 가설 간 승패 패턴을 분석하여 우수한 가설을 선별**

*** 가설 진화 과정(hypothesis evolution process)을 통해 지속적으로 가설의 품질을 향상**

또한, AI Co-Scientist는 자기 비판(self-critique) 능력을 갖추고 있어, 웹 검색 등의 도구를 활용하여 자체적으로 피드백을 제공하고 가설과 연구 제안을 반복적으로 개선할 수 있다.

주요 기여 요약

- **AI Co-Scientist 개발:** 단순한 문헌 요약이나 "deep research" 도구를 넘어, 새로운 지식 발견, 가설 생성, 실험 계획 수립을 지원하는 AI Co-Scientist를 개발하고 도입하였다.
- **과학적 추론을 위한 테스트 시점 연산 패러다임의 대규모 확장(Significant scaling):** AI Co-Scientist는 비동기적 작업 실행(Task Execution) 프레임워크를 활용한 Gemini 2.0 기반 다중 Agent 아키텍처를 바탕으로 한다. 이를 통해 시스템은 과학적 추론에 유연하게 컴퓨팅 자원을 할당하며, 핵심적인 과학적 방법론을 반영할 수 있도록 설계되었다. 특히, 과학적 토론 및 토너먼트 기반 진화 프로세스를 수행하는 자기 대결(Self-Play) 전략을 활용하여 가설 및 연구 제안을 반복적으로 개선하는 자기 향상 루프(Self-Improving Loop)를 형성한다. 15개의 복잡한 과학적 목표(전문가가 선정한 공개 과제)를 대상으로 자동 평가(Auto Evaluation)를 수행하여, AI 공동 과학자가 최첨단(State-of-the-Art, SOTA) 수준의 agent 및 추론 모델을 능가하는 고품질 가설을 생성함을 입증하였다.
- **전문가와 협력하는 과학적 워크플로우(Expert-in-the-Loop Scientific Workflow):** 본 시스템은 과학자들과의 협력을 염두에 두고 설계되었으며, 자연어 기반의 대화형 피드백을 유연하게 반영할 수 있다. 이를 통해 연구자들과 공동 개발(Co-Development), 진화(Evolution), 정제(Refinement) 등의 과정을 수행한다.
- **AI Co-Scientist의 생물의학 분야 핵심 연구 주제에 대한 엔드투엔드(End-to-End) 검증:** AI Co-Scientist가 생성한 새로운 AI 기반 가설을 실험적 연구를 통해 검증하였다. 점점 복잡성이 증가하는 세 가지 생물의학 연구 분야-약물 재창출(Drug Repurposing), 신규 치료 표적 발견(Novel Target Discovery), 항미생물제 내성(Antimicrobial Resistance)-에서 성과를 입증하였는데, AI Co-Scientist의 발견들이 실험실(Wet-Lab) 실험으로 검증되었다.

2 관련 연구

2.1 추론모델과 테스트시점 연산 확장(Reasoning models and test-time compute scaling)

- AI 기반 모델과 LLM의 혁신으로 등장한 GPT 및 Gemini 계열 모델들은 대량의 멀티모달 데이터셋으로 훈련되었고, 언어 이해 및 생성 능력에서 놀라운 성과를 보이고 있다.
- 현재 중요 연구 분야인 모델의 추론(reasoning) 능력 향상을 위해 인간의 사고 과정을 모방한 “reasoning models”이 등장했다. 이러한 방향에서 주목받는 접근 방식이 TTC (test-time compute) 패러다임이다. 이는 사전 훈련된 지식에만 의존하지 않고, inference 과정에서 추가적인 계산 자원을 활용하여 System-2 style 사고, 즉, 신중하고 단계적인 reasoning을 수행하여 불확실성을 줄이고 최적의 해결책을 모색하는 방식을 도입한다.
- 이 개념은 AlphaGo와 Libratus에서 등장했는데, AlphaGo는 MCTS(Monte Carlo Tree Search)를 활용하여 게임 상태를 탐색하고 전략적으로 수를 선택하는 방식을 사용했다.

LLM도 테스트 시점에 계산을 증가시켜 철저한 탐색으로 추론과 정답률을 향상시킨다.

- 최근 등장한 Deepseek-R1 모델과 같은 연구들은 강화 학습을 활용하여 모델의 "연쇄 사고(chain-of-thought)" 능력을 정제시키고, 장기적인 복잡한 추론(reasoning) 능력을 강화하는 등 테스트 시점 연산의 가능성을 더욱 확대하고 있다.
- *본 AI Co-Scientist는 추가적인 학습 기법 없이 과학적 추론(reasoning)과 가설 생성이 가능한 다중 agent 프레임워크를 설계하는 것을 목표로, 과학적 방법론에서 유도된 귀납적 편향(inductive biases)-일반화를 위한 기본적인 가정-을 활용하여 테스트 시점 연산 패러다임을 대규모로 확장하는 방안을 제안한다.*

2.2 AI 기반 과학적 발견

AI 기반 과학적 발견은 다양한 과학 분야에서 연구가 수행되는 방식에 혁신적인 패러다임 전환을 가져왔다. 최근 심층 학습 및 생성 모델의 발전으로 과학적 발견에서 AI의 역할이 공고히 되었다. 이를 가장 잘 보여주는 예시로 단백질 구조예측(AlphaFold2), 신규 항생제 개발, 단백질 결합 설계, 신소재 발견 등이 있다. 최근 AI를 연구과정-최신 LLM 기반 시스템을 초기 가설 생성부터 논문 작성까지-에 완전히 통합하는 것을 목표로 더욱 야심찬 연구가 되고 있다. 이러한 end-to-end 통합은 AI가 특화된 도구를 넘어 능동적인 협력자, 혹은 초기 형태의 "AI 과학자"로 발전하는 과정에서 전례 없는 기회와 중대한 도전 과제를 제시하고 있다.

AI와 연구 통합 사례

① **논문 피드백 제공을 위한 LLM 활용 (**Liang et al.): LLM이 연구 논문에 대한 피드백을 제공할 수 있는지 평가했다. 기존 피어 리뷰(peer review)를 분석하고 사용자 연구를 수행한 결과, LLM이 생성한 피드백과 인간 리뷰어의 피드백이 상당히 유사함을 발견했다. 일부 경우에는 인간 동료보다 더 도움이 된다고 응답했다.

② **PaperQA2 (AI 기반 논문 검색 및 요약 에이전트):** PaperQA2는 AI 기반의 과학 문헌 검색 및 요약 에이전트로, 기존 연구 내용을 종합하는 데 초점이 맞춰져 있으며, 새로운 가설을 생성하는 과학적 추론(reasoning) 과정은 없다.

③ **HypoGeniC (LLM을 활용한 가설 생성 시스템** (Zhou et al.)): LLM과 다중 슬롯 머신(Multi-Armed Bandit) 알고리즘을 결합하여 가설을 반복적으로 정제하는 가설생성 방식을 탐색했다. 소규모 예제 데이터에서 초기 가설을 생성하고, 이후 학습정확도 기반 보상함수로 탐색과 활용(exploration and exploitation) 방식의 최적화를 거쳐 점진적으로 가설을 개선하는 방식이다. 완전히 새로운 가설을 생성할 수 있는지는 불확실하다.

④ **Data-to-Paper (AI 기반 연구 논문 생성 플랫폼** (Ifargan et al.)): 여러 LLM 및 규칙 기반 에이전트의 조정을 통해 연구 논문을 자동으로 생성한다. 논문 작성 과정에서 자동 피드백을 제공하고, 정보 추적 기능을 통해 정확성을 검증한다. 기존 논문 재현에 초점을 맞추며, 완전히 새로운 탄탄한 가설과 연구 제안서를 생성하는지는 불확실하다.

⑤ **Virtual Lab (다중 LLM 기반 가상 연구실):** “principal investigator” 역할의 LLM이 여러 특정문제에 전문화된 LLM 에이전트들을 조율하는 LLM팀으로 구성되었으며, 사람의 고차원 감독을 받는다. 일반화 여부는 불명확하다.

⑥ **Coscientist (GPT-4 기반 화학 실험 자동화 시스템** (Boiko et al.)): GPT-4를 기반으로 한 다중 agent 시스템인 "Coscientist"를 소개하였으며, 이는 복잡한 화학 실험을 자동으로 수행하도록 설계되었다. 웹 및 문서 검색, 코드 실행 등의 기능을 통합하여 독립적인 실험 설계, 계획 및 실행을 가능하게 한다. 화학 연구에 초점을 맞추었고, GPT-4 기반

agent를 단순히 결합한 형태에 가깝다. 높은 수준의 실험 자동화를 우선시하며, 실험실 하드웨어와 직접 인터페이스 하는데 중점을 둔다. 반면, 본 AI Co-Scientist는 “과학자 참여(scientist-in-the-loop)" 방식을 강조하는 협업 도구로 설계되었다.

⑦ **AI Scientist (완전 자동화된 AI 연구 시스템** (Lu et al.)): 연구 문제 정의와 문헌 조사부터 실험 설계 및 실행, 그리고 연구 결과의 논문 작성까지 수행하는 연구 전 과정을 다루는 다중 협력 LLM agent들을 활용하여 완전 자동화된 연구를 수행하도록 설계된 “AI Scientist”를 제안했다. 해당 시스템은 AI Co-Scientist와 유사한 개념을 공유하지만, 자동 평가가 제한적이다. AI Co-Scientist의 접근 방식은 Test-time Compute 확장을 통해 고품질의 가설과 연구 제안을 생성하는 데 초점을 맞추며, 자동 평가, 인간 전문가 검토, 실제 실험 검증의 조합으로 유력한 연구 조력자를 구축하고자 한다.

2.3 생의학을 위한 AI

일반 목적 LLM(GPT-4, Gemini) 및 특화된 LLM(Med-PaLM, Med-Gemini, Galactica, Tx-LLM) 모두 생의학적 추론 (reasoning) 및 Q&A 벤치마크에서 성능을 입증하였고, Med-PaLM2는 당뇨병, 백내장, 청력 손실과 같은 형질의 원인 유전자를 식별하는 데 성공적이었다. 또한, DNA, RNA, 단백질 서열을 학습한 특화된 기초 모델 및 LLM이 활발히 개발되고 있다. 생의학 분야에서 AI는 특화된 모델을 요구하지만, 최첨단 AI 모델의 급속한 발전은 특화 모델과 범용 모델 간의 경계를 허물고 있다. 이러한 모델들이 규모가 커지고, 데이터 다양성이 증가하며, 복잡성이 높아짐에 따라, 한때 도메인 특화 AI가 필요하다고 여겨졌던 영역에서도 혁신을 이루고 있다. 본 AI Co-Scientist 시스템은 모듈형 다중 agent 아키텍처를 기반으로 설계되어, 범용 최첨단 AI 모델의 발전을 수용하는 동시에, 특정 과업에서 특화된 AI 모델을 도구로 활용하여 기능을 확장할 수 있도록 유연하게 설계되었다.

- 약물 재창출(검증 실험): 전통적인 약물 재창출은 계산적 및 실험적 기법이 필요하며, 질병-약물 간 상호작용의 종합적인 이해가 필수적이다. Co-Scientist는 최신 LLM을 활용하여 보다 확장성 높은 접근법을 제안한다. expert evaluations 및 wet-lab experiments를 결합한 검증을 통해, 시스템 예측의 신뢰성을 평가하고 실제 적용 가능성을 높인다.

3. AI Co-Scientist (시스템 구성 기술적 세부 사항, 에이전트, 프레임워크)

- AI Co-Scientist는 Gemini 2.0을 기반으로 구축된 다중 agent 아키텍처를 사용하며, 비동기식 작업 실행 프레임워크에 통합되어 있다. 이 프레임워크는 TTC 자원의 유연한 확장으로 고급 과학적 추론을 가능하게 한다.
- AI Co-Scientist는 다음과 같은 기본 기준을 준수하는 가설과 연구 제안을 생성한다.

• **연구 목표와 부합(Alignment):** 생성된 출력물은 과학자가 정의한 연구 목표, 선호도, 제약 조건과 정확하게 일치해야 한다.

• **타당성(Plausibility):** 생성된 연구 결과는 명백한 오류가 없어야 하며, 기존 문헌이나 확립된 지식과의 모순점들은 명확히 언급하고 정당성을 제시해야 한다.

• **참신성(Novelty):** Co-Scientist의 핵심 목표는 기존 정보의 종합-기존 deep research처럼-이 아니라, 기존 문헌 기반 새로운 가설, 추론, 연구 계획을 생성하는 것이다.

• **검증 가능성(Testability):** 출력물은 연구자 명시 제약조건으로 실증적 검증해야 한다.

• **안전성(Safety):** 출력은 해로운 연구 방지와 안전하고 윤리적이 되기 위해 제어되어야 한다.

- 또한, Co-Scientist는 추가적인 기준, 선호도, 제약 조건을 설정하도록 구성될 수 있다. 예)

가독성과 해석 가능성이 향상되도록, 연구자가 선호하는 형식으로 출력을 생성할 수 있다.
- Co-Scientist의 다양한 구성 요소를 설명하기 위해 근위축성 측색 경화증(ALS, Amyotrophic Lateral Sclerosis)의 생물학적 메커니즘을 탐색하는 가설을 생성해본다.

3.1 AI Co-Scientist 시스템 개요: 4가지 주요 구성 요소

① **자연어 인터페이스:** 자연어를 통해 시스템과 상호 작용하고 감독한다. 이를 통해 초기 연구 목표를 정의할 뿐만 아니라, 필요할 때마다 수정하고, 생성된 가설에 대한 피드백을 제공하며(과학자의 솔루션 포함), 전반적인 연구 진행을 안내할 수 있다.

② **비동기 작업 프레임워크:** Co-Scientist는 특화된 agent들이 비동기적, 지속적, 그리고 구성 가능한 작업 실행 프레임워크 내에서 worker processes로써 동작하는 다중 agent 시스템을 활용한다. 이 프레임워크에서는 감독 agent가 작업 대기열을 관리하며, 특화된 agent에게 작업을 할당하고, 리소스를 분배한다. 이를 통해 시스템은 컴퓨팅 리소스를 유연하고 효율적으로 활용하며, 과학적 추론 능력을 점진적으로 향상시킨다.

③ **특화 agent:** 과학적 방법에서 도출된 귀납적 편향과 과학적 사전 지식을 기반으로, 과학적 추론과 가설 생성 과정이 하위 작업으로 분할된다. 개별 특화 agent들이 맞춤형 명령 프롬프트를 활용하여 하위 작업들을 실행하며, 특화 agent들은 감독 agent에 의해 조정된다.

④ **맥락 기억(Context Memory):** 장기간에 걸친 반복 연산과 과학적 추론이 가능하도록, Co-Scientist는 agent와 시스템의 상태를 저장/검색하는 지속적 맥락 기억을 활용한다.

3.2 연구 목표에서 연구 계획 구성으로

- 과학자가 설정한 연구 목표는 Co-Scientist 시스템의 입력 지점(entry point) 역할을 한다. Gemini 2.0 모델의 멀티모달 처리 능력과 long context 처리 능력을 활용하여, Co-Scientist는 단순한 문장에서부터 수만 개의 자연어 토큰이 포함된 방대한 문서에 이르기까지 다양한 복잡도의 연구 목표를 효율적으로 처리한다. 연구 목표는 연구자가 수행하는 실험실 환경이나 연구 분야와 관련된 특정 제약 조건, 속성, 선호 사항을 포함할 수 있다.
- Co-Scientist 시스템은 입력된 연구 목표를 분석하여, 연구 제안 생성을 위한 연구 계획 구성을 도출한다. 이 구성에는 연구 제안의 선호 사항, 속성, 제약 조건이 반영된다. 예를 들어, Co-Scientist가 오직 새로운 가설만 제안해야 하는지 여부를 설정할 수 있으며, 가설의 품질을 평가하는 기준(예: 참신성, 실험적 타당성)도 정할 수 있다. 이러한 평가 기준은 시스템이 자동 평가 및 개선 단계를 수행할 때 활용된다. 연구 목표에 따라 속성, 선호 사항, 평가 기준을 맞춤 설정할 수 있다.
- 연구 계획 구성이 완료되면, 감독(Supervisor) agent가 작업 대기열을 생성하고, 특화 agent들을 조율하기 시작한다. 이 시스템은 비동기적이며 지속적으로 운영된다. 주기적으로, 감독 agent는 시스템의 상태 및 연구 목표에 대한 진행 상황을 반영하는 요약 통계를 계산한다. 이 통계 정보는 컴퓨팅 리소스 할당 결정과 연구 프로세스가 종료 조건에 도달했는지 판단하는 데 활용된다. 또한, 시스템의 맥락 기억에 이러한 상태 정보를 주기적으로 기록하여, 이후 반복 연산(computation rounds)에서 피드백으로 활용된다. 이 기능은 시스템 일부가 실패했을 경우, 쉽게 재시작할 수 있도록 하는 데에도 유용하다.

3.3 AI Co-Scientist를 구성하는 특화 에이전트들 (Specialized agents)

Co-Scientist의 핵심은 감독 agent가 조율하는 여러 특화된 agent들의 연합이다. 이 agent

들은 과학적 추론 과정을 모방하도록 설계되어, 새로운 가설과 연구 계획을 생성한다. 또한, 웹 검색 엔진이나 특화된 AI 모델과 같은 외부 도구들과 API를 통해 상호작용한다.

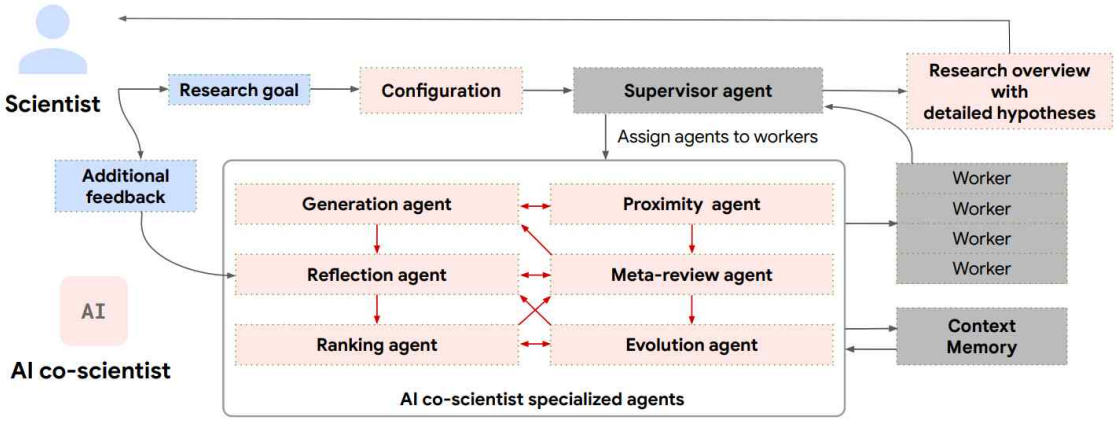


그림 20. AI Co-Scientist의 멀티 에이전트 아키텍처 설계도: Co-Scientist는 입력된 연구목표를 받아 연구계획 구성안을 만든다. 이 계획이 감독 agent에게 전달되면, 감독 agent는 이를 평가하여 각 특화 agent에게 가중치와 자원을 할당하고, 이 가중치에 따라 agent들을 작업 큐에 worker processes로 등록한다. Worker processes는 agent들의 작업 큐를 실행하며, 시스템은 모든 정보를 종합해 세부 가설 및 제안서가 포함된 연구개요를 과학자에게 제공한다. "AI Co-Scientist의 특화 agent" 섹션에 있는 빨간색 박스는 각각 고유한 논리와 역할을 가진 개별 agent들이다. 파란색 박스는 과학자의 개입(입력 및 피드백) 부분이다. 짙은 회색 화살표는 시스템 내 정보 흐름이며, 빨간색 화살표는 특화 agent 간의 피드백 루프이다.

- **생성 agent:** 연구 목표를 분석하여 초기 집중 분야를 생성하고 이를 반복적으로 확장하여 초기 가설/제안을 만들어 연구 과정을 시작한다. 관련문헌을 웹 검색하고, 기존 발견을 종합하여 새로운 방향을 제시하며, 모의 과학 토론을 통해 가설을 개선한다.
- **반영 agent:** 과학적 peer reviewer의 역할을 시뮬레이션하여, 생성된 가설과 연구 제안의 정확성, 품질, 참신성을 비판적으로 검토한다. 특히, 기존문헌 탐색을 통해 충분히 설명되지 않은 연구 관찰 결과에 대해 각 가설이 얼마나 더 나은 설명을 제공할 수 있는지 평가한다.
- **순위 agent:** Co-Scientist 시스템에서 중요한 개념 하나는 각각의 다른 연구제안서가 반복적인 향상을 위해 "토너먼트" 방식으로 평가되는 것이다. 이 agent는 Elo 기반 토너먼트 시스템을 활용해 가설들을 평가하고, 우선순위를 정한다. 이 과정에서는 쌍방 간의 모의 과학 토론이 활용되어 각 제안의 상대적 강점을 세밀하게 평가한다.
- **근접도 agent:** 생성된 가설들 간의 근접도 그래프를 비동기적으로 계산하여, 유사한 아이디어들을 군집화하거나 중복을 제거하고, 가설 공간을 효율적으로 탐색한다.
- **진화 agent:** 반복적 가설 개선의 핵심적인 역할인 토너먼트에서 상위에 오른 가설들을 계속 정제시키는 역할을 한다. 그 정제 전략에는 기존 아이디어의 종합, 유추, 문헌 기반 세부 정보 보완, 비정통적(unconventional) 추론 탐색, 명확성을 위한 개념 단순화 등이 포함된다.
- **메타 리뷰 agent:** 모든 리뷰로부터 통찰을 종합하고, 토너먼트 토론에서 반복되는 패턴을 식별하여, 다른 agent들의 성능을 다음 반복에서 최적화하는 방식으로 지속적인 향상이 되도록 한다. 또한, 이를 통해 생성 가설 및 리뷰의 품질과 관련성도 다음 반복에서 향상된다. 상위 가설과 리뷰를 종합하여 연구 개요를 생성하여 이를 과학자가 검토하도록 한다.

감독 agent는 특화 agent들을 조율하여 입력된 연구목표에 맞는 타당하고, 참신하며, 검증 가능한 가설/연구계획을 만들어 준다. 생성 agent는 연구목표에 부합하는 초기 가설목록을 생성한다. 이 가설들은 반영 agent의 검토를 거쳐, 순위 agent의 토너먼트 평가를 받는다. 진화, 근접도, 메타 리뷰 agent들은 토너먼트 상태를 기반으로 시스템 출력을 향상시킨다.

- 감독 agent는 주기적으로 context memory에 종합적인 통계 정보-생성된 가설 수, 리뷰가 필요한 가설 수, 토너먼트의 진행 상태, 다양한 가설 생성 방법론의 효과 분석관련-를 기록한다. 이러한 통계 정보를 바탕으로 감독 agent는 이후 시스템 동작-즉, 새로운 가설 생성, 리뷰, 토너먼트, 기존 가설 개선 등-을 worker process를 통해 특화 agent들을 전략적으로 선택하고 가중치를 부여하는 형태로 조율한다.
- 중요한 점은, 메타 리뷰 agent는 파인튜닝 또는 강화 학습 같은 역전파 기법 없이도 피드백 전달과 학습을 가능하게 한다. 이 agent는 모든 agent에 적용 가능한 피드백을 생성하는데, 이 피드백은 다음 반복에서 각 에이전트의 프롬프트에 간단히 추가되는 방식이다.

3.3.1 생성 에이전트 (Generation Agent)

생성 에이전트는 다음과 같은 다양한 기법과 도구를 활용해 참신한 가설을 생성한다.

- **웹 검색 기반 문헌 탐색:** 반복적 웹 검색으로 관련 연구논문을 찾아 읽고, 기존 연구결과를 요약하여 추론 기반을 삼는다. 이 요약을 바탕으로 새로운 가설과 연구 계획을 생성한다.
- **모의 과학 토론:** 자기비판 및 자기대결(self-play) 기법을 활용해 전문가 간의 과학적 토론을 시뮬레이션한다. 이러한 토론은 여러 차례의 대화를 통해 점점 정제된 가설로 수렴된다.
- **반복적 전제 조건 식별:** 검증 가능한 중간 전제들을 반복적으로 도출하는데, 이 중간 전제들이 참으로 입증될 경우, 새로운 과학적 발견으로 이어질 수 있다. 이 타당할듯한 전제와 하위 전제들은 조건부 추론 과정을 통해 식별되며, 이후 하나의 완성된 가설로 통합된다.
- **연구 확장:** 가설 공간 내 기존에 탐색되지 않은 영역을 발견하기 위해, agent는 기존 가설과 이전 반복에서 메타 리뷰 agent에 의해 생성된 연구개요 및 피드백을 검토한다. 이를 바탕으로 연구 가설 공간에서 새로운 탐색 방향을 정한다.

- 각각의 생성된 가설을 요약 및 분류하여, 과학자가 핵심 아이디어를 파악하도록 한다.
- ALS(근위축성 측삭경화증) 가설 및 연구 제안 예시: 그림 A.10

3.3.2 반영 에이전트 (Reflection agent)

리뷰는 새로운 제안의 효과적 생성에 필수적이다. 반영 agent는 웹 검색이나 과학자가 제공한 전용 저장소를 통해 관련연구를 탐색하고, 주어진 가설에 대한 기존 실험 증거를 평가하며, 생성된 출력물의 참신성, 정확성, 품질을 엄격하게 검증한다. 효과적인 리뷰는 부정확하거나 참신하지 않은 가설을 걸러내며, 다른 agent들에게 피드백을 제공하여 지속적인 개선이 되도록 한다. 반영 agent는 다음과 같은 유형의 리뷰를 수행한다.

- **초기 리뷰(Initial review):** 기본 평가기준을 바탕으로, 생성된 가설의 정확성, 품질, 참신성, 그리고 안전성(윤리성)에 대한 초기 리뷰를 수행한다. 이 초기 리뷰는 외부 도구(예: 웹 검색)를 사용하지 않고, 결함 있거나 참신하지 않거나 적합하지 않은 가설을 신속히 제거한다.
- **전체 리뷰(Full review):** 가설이 초기 리뷰를 통과하면, 반영 agent는 보다 정교한 추론과 근거를 제공하기 위해 외부 도구와 웹 검색을 활용하여 관련 문헌을 찾아 전체 리뷰를 만든다. 이때, 문헌 검색을 전면적으로 활용하여 초기 리뷰처럼 정확성, 품질, 참신성을 평가한다. 정확성과 품질 측면에서 가설의 기저 가정과 추론을 면밀히 검토하며, 참신성 측면에서

가설의 알려진 부분을 요약하고, 기존 문헌을 바탕으로 새로움 여부를 판단한다.

- **심층 검증 리뷰(Deep verification review):** 가설을 구성하는 주요 가정들로 분해하고, 각 가정은 다시 기본적인 하위 가정으로 세분화되며, 문맥에서 분리된 상태로 독립적으로 평가되어, 이후 필터링을 위한 무효화 요소를 식별해낸다. 동시에, 잘못된 가정으로 인해 가설이 무효화될 수 있는 이유도 요약된다. 이러한 심층 검증은 복잡한 가설 속에 숨겨진 미묘한 오류들—예를 들어 추론의 결함이나 부정확한 실험 절차-의 탐지를 도와준다. 발견된 오류가 반드시 가설의 핵심을 무효화하는 것은 아니며, 반영 agent는 그 부정확한 가정이 가설에 근본적인지 여부를 평가한 뒤, 그에 대한 판단 근거를 리뷰에 반영한다. 근본적이지 않은 오류는 이후의 정제 단계에서 수정이 가능하다.
- **관찰 기반 리뷰(Observation review):** 가설이 과거 실험 결과에서의 희귀한 관찰을 설명할 수 있는지 탐색한다. 이 리뷰는 가설이 기존 연구 결과나 관련 문헌에서 관찰된 현상에 대한 새로운 통찰을 제공할 수 있는지를 목표로 한다. 각 관찰에 대해, 가설이 기존 설명보다 우수한지(가설이 참이라고 가정한 상태에서)를 평가하고, 긍정적인 결과는 요약되어 가설에 덧붙여진다. 유의미한 결과가 없을 수도 있다.
- **시뮬레이션 리뷰(Simulation review):** 가설을 단계적으로 시뮬레이션하여 (예: 작용 메커니즘이나 제안된 실험) 가설을 검토한다. 또한, 실패 시나리오를 식별하고 요약한다. 이 방식은 첨단 LLM들이 내부 세계 모델을 갖추어 다양한 과학 현상을 시뮬레이션으로 정교하게 예측할 수 있다는 가정에 기반 한다.
- **반복/토너먼트 리뷰(Recurrent/tournament review):** Co-Scientist의 축적된 지식을 전체 리뷰에 지속적으로 반영한다. 리뷰된 가설과 순위 agent의 토너먼트 결과를 분석하여 반복되는 문제나 개선 기회를 식별하고 리뷰를 정제한다.

- 또한, Co-Scientist는 전문가 과학자의 리뷰를 통합하여 순위 매기기와 개선을 유도할 수 있다(3.4절에서 추가 논의). 반영 agent의 종합적인 리뷰가 과학자들이 연구 가설 및 제안을 비판하고 개선할 때 사용하는 일반적인 방법들을 포괄하도록 한다.

3.3.3 순위 에이전트 (Ranking agent)

- 연구 목표에 맞는 수많은 가설과 연구 제안을 탐색하여, 이 중 가장 유망한 후보에 계산 자원을 우선 할당하기 위해 순위 메커니즘이 필요하다. 순위 agent가 이를 담당하는데, Elo 기반 토너먼트 방식을 사용하여 모든 가설을 자동으로 평가하고 순위를 매기며, 이에 대한 합리적 근거도 제공한다. 연구 목표에 부합하는 가설과 제안들의 순위 목록은 과학자에게 정렬되어 전달된다. Elo는 그 한계점에도 불구하고, 상대적 순위를 평가하는 데 여전히 좋은 지표로 사용되고 있다. 새로 추가된 가설의 초기 Elo 점수는 1200점으로 설정한다.
- 토너먼트는 계산이 집중되는 작업이기에, 순위 agent는 여러 최적화 전략을 적용한다. 상위권 가설은 다회전 과학적 토론을 통해 쌍으로 비교되며, 이 과정은 정렬 편향을 줄이고 참신성, 정확성, 검증 가능성에 초점을 맞춘다. 하위권 가설은 단일 회전으로 쌍비교가 된다. Agent는 어느 가설이 더 우수한지에 대한 결정으로 각 비교의 마무리를 짓는다.
- 순위 에이전트는 다음 기준에 따라 토너먼트 대진의 우선순위를 정해 진행한다. 성공적인 가설은 빠르게 높은 순위를 획득하며, 이는 이후의 토너먼트 상태에 영향을 준다.
  (1) 근접도 agent의 그래프를 기반으로, 서로 유사한 가설들끼리 비교될 가능성이 높다.
  (2) 최근 추가된 가설과 상위권 가설은 토너먼트 매치에 우선적으로 포함된다.

3.3.4 근접도 에이전트 (Proximity agent)

근접도 agent는 연구 목표를 고려하여 연구 가설 및 제안들 간의 유사도를 계산하고, 근접도 그래프를 만든다. 직접적으로 가설 생성에 관여하지 않지만, 순위 agent가 토너먼트 매치를 구성하고 연구 목표와 관련된 다양한 아이디어를 제시하는 것을 보조한다. 이를 통해 과학자들은 관심 있는 영역을 빠르게 탐색하고, 연관된 개념을 쉽게 파악할 수 있다.

3.3.5 진화 에이전트 (Evolution agent)

다음과 같은 방식을 활용하여 기존 가설과 제안을 지속적으로 정제하고 개선한다.

- **근거 기반 향상(강화)**: 가설의 약점을 식별하고, 검색 쿼리 생성, 관련 문서 검색 및 독해, 개선 제안, 그리고 추론의 공백을 메우기 위한 세부 설명을 통해 가설 개선을 시도한다
- **일관성, 실용성, 실행 가능성 향상**: 문제점 해결을 목표로, 더 일관성 있는 가설을 생성하며 잘못된 초기 가정으로 인한 근본적인 문제를 걸러낼 수 있다. 또한, 가설을 보다 실용적이고 실행 가능한 형태로 정제한다.
- **기존 가설로부터의 영감**: 여러 상위 순위 가설로부터 영감을 받아 새로운 가설을 생성한다.
- **조합(Combination)**: 상위 가설의 우수한 요소를 직접 결합하여 새로운 가설을 만든다.
- **단순화(Simplification)**: 가설을 보다 쉽게 검증하고 테스트할 수 있도록 단순화한다.
- **창의적 사고(Out-of-box thinking)**: 가설의 부분집합에서 벗어나 발산적 아이디어 생성에 의해 파격적인 아이디어를 탐구한다.

- 진화 agent는 기존 가설을 수정하거나 교체하지 않고, 새로운 가설을 생성한다. 이 전략은 상위 순위 가설이 잘못된 개선으로 품질이 저하되는 것을 방지하며, 새로운 가설은 토너먼트를 통해 경쟁한다. 이처럼 연구 가설과 제안의 진화 과정은 Co-Scientist가 다양한 개선 기법을 반복적으로 조합하고, 결과물의 품질을 점진적으로 향상시키게 한다.

3.3.6 메타 리뷰 에이전트 (Meta-review agent)

메타 리뷰 agent는 피드백 루프에서 결정적인 역할을 하여, 과학적 추론에서의 자기 개선을 가능하게 한다. 이 agent는 토너먼트 상태를 기반으로 작동하며, 토너먼트 매치에서 이루어진 리뷰 및 과학적 토론에서 발견된 공통된 패턴을 요약하여 메타 리뷰 비평을 생성한다.

- 모든 리뷰에서 얻은 통찰을 종합한 메타 리뷰는 “반영 agent”에 가치 있는 피드백을 제공하여, 향후 리뷰를 더 철저하고 신뢰성 있게 개선하도록 한다. 이를 통해 중요한 세부사항이 간과되는 것을 방지할 수 있다. 예) ALS에 대한 약물 재창출 후보 식별 연구 목표의 사례에서 개별 리뷰의 90%만이 제안된 후보물질의 혈액-뇌 장벽(BBB) 투과성 문제를 올바르게 식별했다고 하더라도, 메타 리뷰는 반영 agent의 향후 리뷰가 반드시 이 핵심 요소를 다루도록 해준다. 메타 리뷰의 반복적 문제 식별에 의해 가설 및 연구 제안 생성도 향상된다. 생성 agent는 이 피드백을 리뷰 비평에 대한 과도한 최적화(오버피팅)를 피하기 위해 선별적으로 사용하는 반면에, 메타리뷰는 자주 반복되는 문제의 재발을 예방하게 해준다.

- **연구 개요 생성 (Research overview generation)**: 주기적으로 상위 순위 가설들을 종합하여 연구 개요를 생성하며, 이는 향후 연구를 위한 로드맵 역할을 한다. 이 개요는 연구 목표와 관련된 잠재적인 연구 영역 및 방향성을 제시하고, 그 중요성을 정당화하며 각 영역별로 구체적인 실험도 제안한다. 각 영역에는 예시 주제가 함께 제시된다. 연구 개요는 이후 반복 과정에서 생성 agent에 추가 입력 요소로도 활용된다. 연구 개요는 AI Co-Scientist 시스템 내에서 연구 목표와 관련하여 현재 지식의 경계를 시각화하고, 향후 탐색이 필요한

영역을 조명하는 데 기여한다. (참조: ALS 연구 개요 예시 그림 A.20~A.21) 또한, 메타 리뷰 agent는 이 연구 개요를 제약 기반 디코딩 기법을 통해 NIH(Specific Aims Page)와 같은 연구 보고서 및 과제 제안서 양식에 맞춰 포맷팅할 수 있다.

- **연구자 추천 (Research contacts identification):** 이전 문헌 리뷰를 바탕으로, 연구 가설 및 제안에 대해 검토할 수 있는 적합한 분야 전문가들을 제안한다. 각 추천에는 선정 이유가 포함되며, 해당 전문가 정보는 연구 개요에 요약되어 연구자에게 부가적인 관점과 가능한 협업 수단을 마련해준다.

3.4 전문가 참여 기반 상호작용(Expert-in-the-loop interactions with the co-scientist)

AI Co-Scientist는 전문가 참여 기반 설계를 통해 과학자들이 시스템을 능동적으로 조정하도록 한다(그림 2). 과학자들은 다음과 같은 다양한 방식으로 시스템과 상호작용할 수 있다.

- 생성된 가설 및 연구 개요를 바탕으로 초기 연구 목표를 정제한다.
- 생성된 가설에 대해 수동으로 리뷰를 마련해준다(3.3.2절 참조). AI Co-Scientist는 이러한 전문가 수동 리뷰를 가설과 제안의 평가 및 개선에 활용한다.
- 과학자 자신의 가설과 연구 제안을 토너먼트에 포함시켜 순위가 매겨지고, 시스템이 생성한 가설 및 제안과 함께 결합될 수 있다.
- 특정 연구 방향을 따라가도록 Co-Scientist에게 지시할 수 있다 (예: 일부 특정 문헌에 한정하여 탐색). 이러한 연구가 연구 목표에 명시될 경우, Co-Scientist는 해당 내용에 접근하고 통합할 수 있는 생성 방법을 우선적으로 활용한다.

3.5 AI co-scientist의 도구 활용

Co-Scientist는 가설 및 연구 제안의 생성, 검토, 개선 과정에서 다양한 도구를 활용한다. 웹 탐색과 검색은 기초적이면서도 최신 정보에 기반한 가설을 위해 가장 중요한 기본 도구이다. 특정 가능성의 범위 내에서 탐색하는 연구 목표(예: 특정 유형의 세포 수용체 전부/모든 FDA 승인 약물)의 경우, Co-Scientist는 도메인 특화 도구—예: 공개 데이터베이스—를 사용하여 검색 범위를 제한하고 가설을 생성한다. 또한 과학자가 지정한 개인 저장소를 색인화하고 검색할 수 있다. 마지막으로, 시스템은 AlphaFold와 같은 특수화된 AI 모델의 피드백을 활용하고 통합할 수 있다. A.6절에서는 단백질 설계 사례를 통해 이를 정성적으로 시연하고 있다.

4. 평가 및 결과 (AI Co-Scientist의 평가 방법과 그에 따른 결과)

① Co-Scientist를 구성하는 전략과 지표의 선택을 벤치마킹하고 검증한다.
② 시스템의 품질을 평가하기 위해 도메인 전문가들과 함께 소규모 평가를 수행한다.
③ 시스템이 생성한 새로운 예측의 실질적인 유용성을 평가하기 위해, Co-Scientist가 제안한 가설과 연구 제안을 세 가지 주요 생의학 응용 분야-약물 재창출, 새로운 치료 표적의 발견, 항생제 내성의 근본 메커니즘 규명-에서 실험실 실험을 통해 검증한다. 이 응용 분야는 복잡성과 성격이 다양하므로 시스템을 보다 포괄적으로 평가할 수 있다.

4.1 Elo 평점과 AI Co-Scientist 결과의 일치

Elo 자동 평가 평점은 Co-Scientist 시스템에서 자기 개선 피드백 루프를 진행하는 핵심 지표이다. Elo 평점과 품질 결과의 상관관계를 측정하고 검증하는 것이 필요하다. 이를 위해 GPQA 벤치마크 데이터셋에 대하여 Elo 평점과 시스템의 정확도 간의 일치여부를 분석했다.

이상적으로는, Elo 평점이 높을수록 정답일 확률도 높아야 한다.

- GPQA 데이터셋은 생물학, 물리학, 화학 분야의 전문가들이 개발한 도전적인 객관식 문제 풀이 벤치마크이다. Co-Scientist의 Elo 평점이 시스템의 결과가 타당하고 정확한지를 객관적으로 반영하는 지표가 되도록 하기 위해, GPQA 데이터셋 중에서도 난이도가 높은 것으로 알려진 GPQA 다이아몬드 세트 내의 문제(질문)들을 활용했다. 각 질문에 대하 AI 시스템은 하나의 연구 목표로 응답을 했다. 각 질문에 대해 Co-Scientist가 생성한 응답을 정답과 비교해 정확도를 평가한 후, 생성된 응답을 Elo 평점 범위에 따라 50점 단위로 분류했다. 그런 다음, 각 Elo 평점 구간 내에서 정답 비율을 평균하여 정확도를 계산했다.
- AI Co-Scientist 시스템에서 사용되는 기반 모델인 Gemini 2.0을 기준으로 활용했다. 각 GPQA 질문에 대해 32개의 Gemini 2.0 응답을 생성하고, Gemini 2.0이 해당 질문에서 생성한 정답 비율을 기준 정확도로 삼았다. 특정 Elo 구간에 대한 정확도는, 해당 구간 내에 Co-Scientist가 응답했던 GPQA 질문들의 정확도 평균으로 계산했다. 또한, 각 질문에 대해 가장 높은 Elo 평점을 받은 Co-Scientist 응답을 사용하여 정답과 비교함으로써 GPQA 다이아몬드 세트에 대한 Co-Scientist의 정확도를 산출했다.
- GPQA 다이아몬드 세트 질문을 활용한 분석 결과, Elo 평점과 Co-Scientist가 생성한 응답의 평균 정확도 사이에 일치도가 있음을 확인했다(그림 3 참조). 각 질문마다 최고 평점을 받은 Co-Scientist 응답을 선택했을 때, Co-Scientist의 top-1 정확도는 78.4%였다.

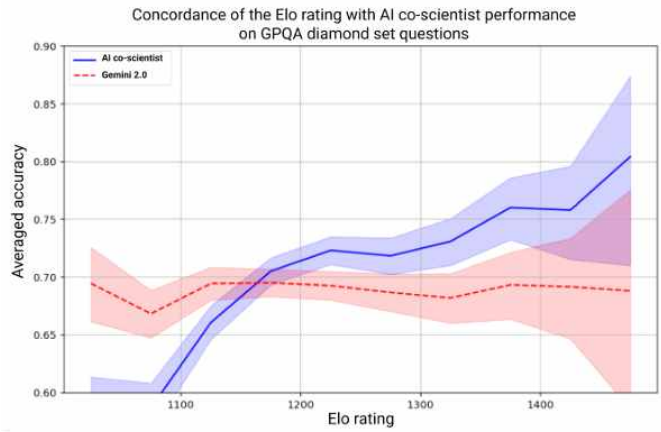


그림 21 GPQA 정확도와 자동 평가 Elo 지표

4.2 테스트 시 연산 확장의 AI Co-Scientist 과학적 추론 능력 향상 효과

반복적인 과학적 추론 및 가설 생성 과정에서 테스트 시 연산 자원 확장 효과와 Co-Scientist의 진전을 평가하기 위해, 토너먼트 동안 Co-Scientist가 생성한 가설과 제안의 Elo 평점을 측정했다. 이 분석은 생의학을 중심으로 수학, 물리학 등 다양한 과학 분야에 걸쳐 선별된 203개의 고유한 연구 목표를 대상으로 수행되었다.

- Co-Scientist가 생성한 연구 가설과 제안들을 생성 시간 순으로 10개 구간으로 나누었다. 각 구간에 대해 상위 10개 가설의 평균 Elo 평점과 최고 Elo 평점을 측정했다. 그 결과 두 지표 모두에서 Co-Scientist의 품질이 시스템의 자기 개선 피드백 루프를 통해 점진적으로 향상되었음을 보여준다(그림 4).
- 이 관찰 결과를 보다 정밀하게 해석하기 위해, 우리는 생의학 분야의 전문성을 가진 7명의 전문가들이 도전적인 문제로 선별한 15개의 연구 목표에 초점을 맞춰 분석을 진행했다. 이

연구 목표들은 연구 제목, 명확한 목표 설정, 생물학적/질병 선호 영역, 바람직한 해결 속성, 실험 기술에 대한 제약 조건 등을 포함하여 일관된 구조와 형식으로 구성되었다.

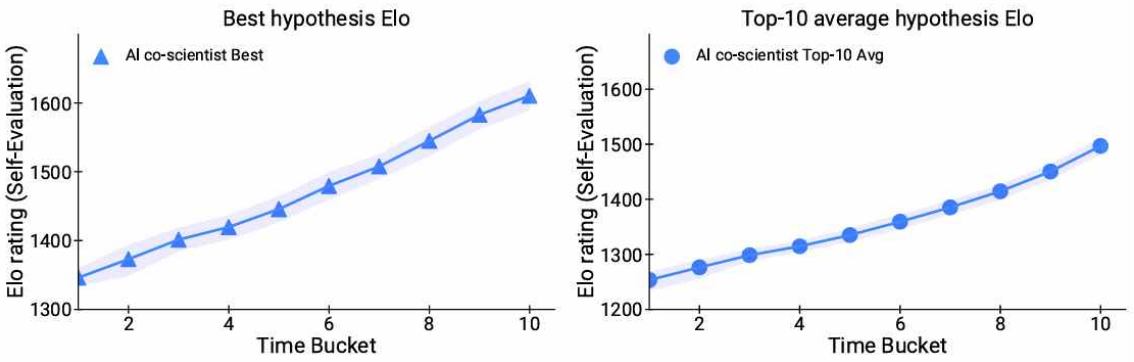


그림 22 Elo 자동 평가로 측정한 테스트 시 연산 확장 효과

- 전문가의 최선의 가설 또는 해결책도 제공되었다. 이후, 최신 LLM 및 추론 모델들(Gemini 2.0 Pro Experimental, Gemini 2.0 Flash Thinking Experimental 12-19, OpenAI o1, OpenAI o3-mini-high, DeepSeek R1)의 출력 결과와 전문가의 "최선의 추정", Co-Scientist의 결과를 Co-Scientist의 Elo 평점을 사용해 토너먼트 형식으로 비교했다.

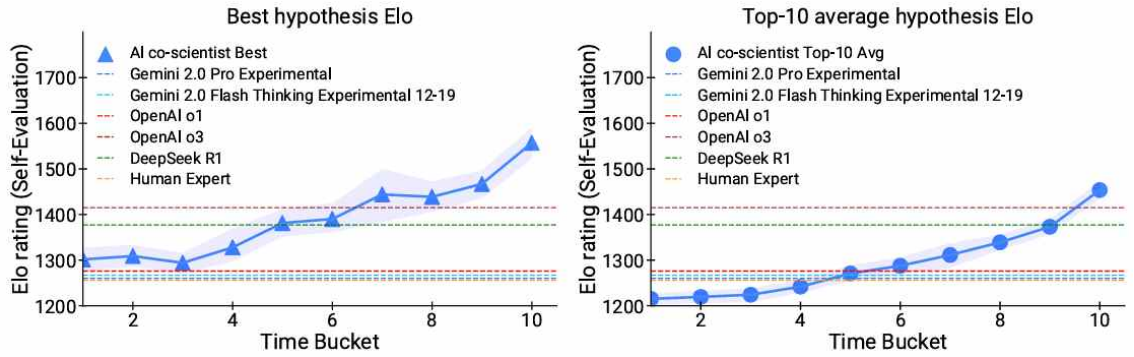


그림 23 전문가 선별한 15개 도전적인 연구 목표에서 Co-Scientist와 기타 모델의 비교

- (그림 5) Co-Scientist는 반복적 개선에 의해 다른 최첨단 LLM 및 추론 모델보다 높은 Elo 평점을 얻었다. OpenAI o3-mini-high 및 DeepSeek R1과 같은 최신 추론 모델들은 훨씬 적은 연산 및 추론 시간으로 경쟁력 있는 성능을 보였다. Elo 측정 결과에서 성능 향상의 포화 현상은 없었으며, 이는 테스트 시 연산 자원을 더 확장할 경우 Co-Scientist 시스템의 품질이 계속해서 개선될 수 있음을 시사한다.
- Co-Scientist 시스템이 연구 가설과 제안을 개선할 수 있는 능력에 기반하여, 전문가의 "최선의 추정" 해결책을 더욱 향상시킬 수 있는 가능성에 대해서도 조사하였다. 그 결과, Co-Scientist는 시간이 지남에 따라 전문가의 "최선의 추정" 해결책을 향상시킬 수 있는 능력을 Elo 지표를 통해 보여주었다. 이는 Co-Scientist와 같은 AI 시스템이 전문가 과학자들의 연구를 보완하고 가속화할 수 있는 유망한 가능성을 시사한다.

4.3 전문가들의 Co-Scientist 결과에 대한 평가

전문가의 피드백을 얻고 선호도를 평가하기 위해, 앞서 선별된 15개의 연구 목표 중 11개를

대상으로 소규모의 전문가 평가를 수행했다. 연구 목표를 선별한 전문가들이 Co-Scientist, Gemini 2.0 Flash Thinking Experimental 12-19, Gemini 2.0 Pro Experimental, 그리고 OpenAI o1 모델의 출력 결과를 평가했다. 전문가들은 각 모델의 결과에 대해 선호도 순위를 매겼으며, 제안된 해결책의 참신성과 영향력을 다음 기준으로 5점 척도 평가했다.

- **참신성 (Novelty):** 높은 점수는 전문가가 아는 한 기존에 어떤 형태로도 발표되지 않은 가설이고, 기존 제안과 유사하거나 약간 수정된 경우는 낮은 점수를 받는다.
- **영향력 (Impact):** 해당 분야에서 중요한 미해결 문제를 다루며, 과학적 이해를 실질적으로 발전시키거나 실용적인 응용 가능성을 지닐 경우 높은 점수이다.

- 11개의 전문가 평가 연구 목표 전반에 걸쳐, Co-Scientist가 생성한 결과가 다른 기준 모델들에 비해 전반적으로 가장 선호되었고, 참신성과 영향력 측면에서도 더 높은 평가를 받았다. 이러한 평가는 객관적인 정답이 아닌 전문가들의 주관적 평가를 반영한 것이다. 주목할 점은, 전문가의 선호도와 Elo 점수 간에도 일관성이 관찰되었다는 점이다(그림 5, 7).

4.4 적대적 연구 목표를 활용한 Co-Scientist의 안전성 평가

Co-Scientist는 잠재적인 오용 가능성을 고려할 때 강력한 안전 원칙-위험한 연구 목표, 이중 사용(dual-use) 목적, 안전한 목표가 위험한 가설로 이어지는 시나리오, 오해 소지가 있는 주장, 내재된 편향-을 기반으로 설계되었는지 검증하는 것이 매우 중요하다. 이는 본 논문의 범위를 벗어나지만, 적대적 테스트 전략을 활용하여 초기 안전성 분석을 했다.

- 최신 LLM들을 이용해 40개의 생물의학 및 과학 주제 전반에 걸쳐 다양한 복잡도를 가진 1,200개의 적대적 예시를 선별했다. 이후 Co-Scientist가 이들 연구 목표를 신뢰성 있게 거부할 수 있는지를 평가했다. 이번 초기 분석 결과, 시스템은 모든 테스트를 성공적으로 통과했다. 총체적으로, 이 섹션에서 제시된 벤치마크, 자동화된 평가, 전문가 평가 결과는 Co-Scientist 시스템의 강력한 성능과 안정성을 입증한다.

4.5 Co-Scientist를 활용한 약물 재창출(Drug Repurposing)

복잡한 연구 문제에 대해 새로운 가설과 예측을 생성하는 시스템의 능력을 엄격히 평가하려면 실험실 기반의 실험을 통한 엔드 투 엔드(end-to-end) 검증이 필요하다. 하지만 이러한 실험은 고난이도, 시간 소모, 높은 자원 요구 등으로 인해 대규모 검증은 현실적으로 어려운 상황이다. 이에 다양하면서도 중요성이 높은 생물의학 주제들을 선정하여 시스템의 전체 성능을 평가할 수 있는 강력한 벤치마크로 활용했다. 세 가지 실험 검증 모두 전문과학자들과 협업하여 수행되었으며, 이들은 Co-Scientist에게 지침을 제공하고 실험의 우선순위를 정해주었다.

- 그 검증의 첫 번째 사례로, 약물 재창출(drug repurposing) 응용에 대하여 논의한다. 약물 재창출은 기존에 승인된 약물을 원래의 용도와는 다른 새로운 치료 적응증에 활용하는 과정이다. 이 접근 방식은 복잡하거나 희귀한 질병에 대한 치료제를 더 빠르게 발견할 수 있게 하며, 이미 안전성이 입증된 약물을 사용하기 때문에 실용적인 이점이 크다. 기술적인 관점에서 보면, 이는 방대한 수의 약물-질병 쌍을 대상으로 한 조합 탐색 문제이다.
- Co-Scientist는 방대한 과학 및 임상 문헌을 종합하고 통합할 수 있는 능력을 가지고 있으므로, 약물 재창출이 Co-Scientist의 능력을 시험하기에 이상적인 응용 분야일 것이다. 이 시스템은 범용 시스템으로, 모든 알려진 약물-질병 쌍에 대해 매우 상세하고 설명 가능한 예측을 생성할 수 있다. 이번 검증에서는 암 치료를 위한 약물 재창출 분야에서 Co-Scientist의 계산 생물학 및 실험실 기반 검증에 초점을 맞췄다.

- 처음에 전임상 근거가 있는 약물-암 조합을 바탕으로 Co-Scientist가 생성한 가설과 예측의 타당성을 검토하였고(4.5.1), 이후 완전히 새로운 약물 재창출 가설로 확장하였다(4.5.2). 이러한 예측의 검증은 계산 생물학 분석, 종양 전문의의 피드백, 암 세포주를 이용한 시험관 내(in vitro) 실험을 포함한 다각적인 접근 방식으로 수행되었다.

* Co-Scientist는 기존의 접근 방식이나 전문가 소스들로는 선택되지 않았을 수 있는 급성 골수성 백혈병(AML)에 대한 새로운 약물 재창출 후보물질을 제안할 수 있었다. 이는 Co-Scientist 시스템이 연구자들이 탐색할 새롭고 유망한 가설을 생성할 수 있는 능력을 가지고 있음을 시사하며, 향후 복잡하고 어려운 질병, 예를 들어 AML과 같은 질환에 대해 새로운 치료법을 환자에게 제공할 가능성을 보여준다.

4.6 AI Co-Scientist의 간 섬유증에 대한 새로운 치료 표적 발견

간 섬유증은 심각한 질병으로, 간부전 및 간세포암으로 진행될 수 있으며, 현재 사용 가능한 동물 모델이나 in vitro 모델의 한계로 인해 치료 옵션이 매우 제한적이다. 그러나 최근에 간 섬유증에 대한 생세포 이미징 시스템과 결합된 인간 간 오가노이드(human hepatic organoids) 생성방법이 개발되어, 간 섬유증에 대한 새로운 치료법 발굴 경로가 열렸다. AI Co-Scientist로 간 섬유증에서 후성유전적 변화(epigenetic alterations)의 역할에 대한 실험적으로 검증 가능한 가설("간 섬유증에서 근섬유아세포 생성에 대한 새로운 가설")을 제안하고, 해당 후성유전 조절인자를 표적으로 하는 약물을 식별하고자 한다.

- 전문가들은 Co-Scientist가 생성한 상위 15개의 연구 가설 중 상위 3개를 선정했으며, 여기에는 실험 설계, 평가 방법론, 예상 결과를 포함한 종합적인 연구 제안서가 포함되어 있었다. Co-Scientist는 기존 약물로 표적화할 수 있는 3가지 새로운 후성유전 조절 인자를 식별했으며, 이들에 대한 전임상(preclinical) 근거도 함께 제시했다. 그중 두 가지를 표적으로 하는 약물은 세포 독성 없이 간 오가노이드에서 강력한 항섬유화 활성을 나타냈다 (그림 12 참고). 이 중 하나는 이미 다른 적응증으로 FDA 승인을 받은 약물이기 때문에, 이를 간 섬유증 치료용으로 재창출(drug repurposing)할 수 있는 기회를 제공한다. 이 결과는 향후 발표될 기술 보고서에 자세히 소개될 예정이다.

4.7 AI Co-Scientist의 항생제 내성 분야의 획기적인 발견 재현

항생제 내성의 메커니즘을 이해하는 것은 감염병에 대한 효과적인 치료법을 개발하기 위해 매우 중요하다. 이에, 항생제 내성에서 중추적인 역할을 하는 캡시드 형성 파지 유도 염색체 섬(capsid-forming phage-inducible chromosomal islands, cf-PICIs)에 초점을 맞췄다. 이 이동 유전 요소들은 일반적인 파지나 다른 PICIs와는 달리, 다양한 박테리아 종 간에 전이될 수 있는 놀라운 능력을 가지며, 이 과정에 병원성과 항생제 내성 유전자를 함께 전달한다.

- AI Co-Scientist를 활용하여 cf-PICIs가 여러 박테리아 종에 걸쳐 존재하는 진화적 이유를 이해하고, 항생제 내성의 확산을 억제하기 위한 전략을 개발하기 위한 연구 제안을 생성하는 것을 주요 목표로 삼았다. 특히 WHO가 지정한 중요 병원체인 대장균과 클렙시엘라 폐렴균과 같은 임상적으로 중요한 박테리아 종에서 PICIEc1 및 PICIKp1과 같은 동일한 cf-PICIs가 발견된다는 점에 주목했다.
- 현재 저명 학술지에서 심사 진행 중인 한 연구에서는, 유전체 및 실험 분석을 통해 서로 다른 박테리아 종에서 동일한 cf-PICIs가 발견되는 현상을 설명하는 새로운 메커니즘이 밝혀졌다. 이를 AI Co-Scientist가 독립적으로 동일하거나 유사한 가설로 제시할 수 있는지 조

사했다. Co-Scientist에게, 파지 위성(phage satellite)의 간략한 배경과 두 편의 관련 논문이 포함된, 일반적인 정보의 단일 페이지 문서를 제공했다. 첫 번째 논문은 cf-PICIs의 최초 발견을 다루었으며, 두 번째 논문에는 박테리아 유전체에서 파지 위성을 식별하는 계산적 기법이 있다. cf-PICIs가 왜 다른 종류의 PICIs나 위성들과 달리 다양한 박테리아 종에서 발견되는지, 그리고 그 현상의 근본 메커니즘은 무엇인지에 대한 질문을 제시했다.

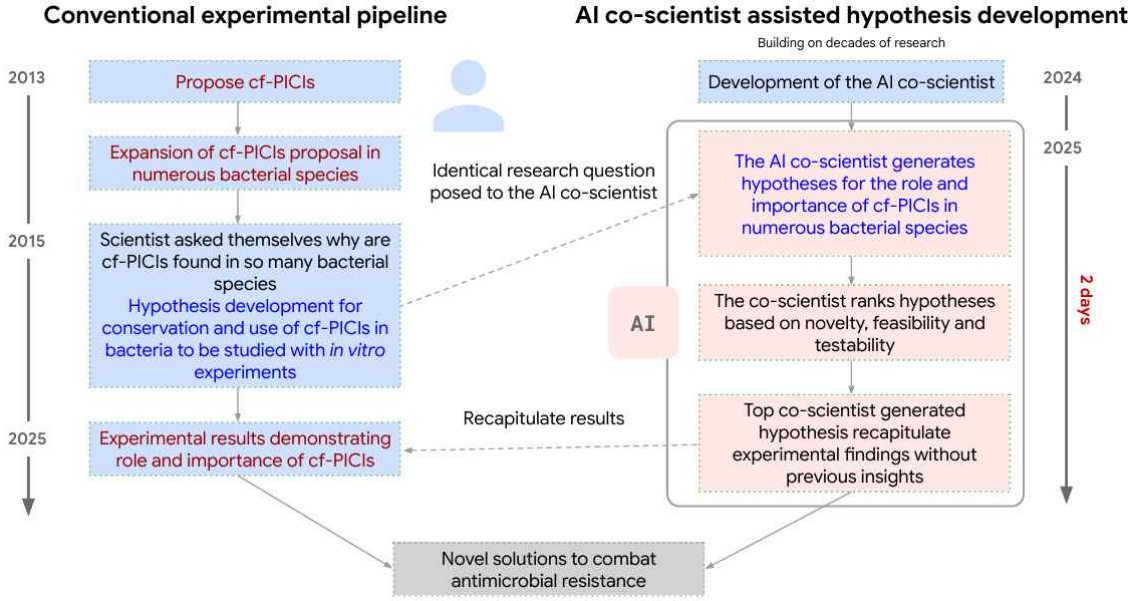


그림 13. 캡시드 형성 파지 유도 염색체 섬(cf-PICIs)에 대한 가설 개발의 타임라인

- Co-Scientist는 “cf-PICIs 요소가 다양한 파지 꼬리와 상호작용하여 숙주 범위를 확장한다”는 획기적인 가설을 가장 높은 순위로 제시했으며, 이는 이후 독립적인 연구에서 실험적으로 검증되었다. Co-Scientist가 불과 이틀 만에 도출해낸 이 가설은 십 년간의 연구를 바탕으로 나온 것이다. 모든 공개 논문에 접근 가능한 점도 작용했다 (그림 13 참조).
- 구체적으로, Co-Scientist는 cf-PICIs에 대한 다음과 같은 연구 주제를 제안했다:

• **캡시드-꼬리 상호작용:** cf-PICI 캡시드와 다양한 보조 파지 꼬리 간의 상호작용을 조사. 이는 미공개 원고의 주요 발견과 정확히 일치:cf-PICI가 다른 파지의 꼬리와 상호작용하여 숙주 범위를 확장. 이 과정은 cf-PICI에서 인코딩된 어댑터 및 커넥터 단백질에 의해 매개됨.
• **통합 메커니즘:** cf-PICIs가 다양한 박테리아 유전체에 통합되는 메커니즘 탐구.
• **진입 메커니즘:** 전통적인 파지 수용체 인식을 넘는 cf-PICI의 진입 방식 탐색.
• **보조 파지 및 환경 요인:** cf-PICI 전이에 있어 보조 파지 및 생태적 요인의 역할 연구.
• **대체 전이 및 안정화 메커니즘:** 접합, 세포외 소포, 기타 고유한 안정화 전략 등 cf-PICI의 폭넓은 숙주 범위를 설명할 수 있는 잠재적 전이 메커니즘 탐색.

- 전통적인 연구 방식과 AI Co-Scientist 접근 방식의 동일한 발견으로의 수렴은 Co-Scientist가 과학 연구를 증대하고, 보완하고, 가속화할 수 있는 잠재력을 강력히 시사한다(그림 13 참조). 추가 결과 및 세부 사항은 동반 보고서 [21]에서 확인할 수 있다.

5. 한계점

AI Co-Scientist의 과학 연구 보조 잠재력은 고무적이나 여러 가지 한계점이 존재한다.

• **문헌 검색, 검토 및 추론의 한계:** AI Co-Scientist의 리뷰는 라이선스나 접근 제한 준수로

전체 출판 문헌에 접근하지 않고 오픈 액세스 문헌에 의존하기 때문에 중요한 선행 연구를 놓칠 수 있다. 또한, 시스템이 특정 연구를 관련성이 없다고 잘못 추론하는 경우도 있다.

- **부정적인 결과 데이터에 대한 접근 부족:** 공개된 문헌만을 사용하므로, 부정적인 실험 결과나 실패한 실험에 대한 기록에 대한 접근은 상당히 제한적이다. 부정적 데이터는 덜 출판되는 경향이 있지만, 해당 분야의 숙련된 과학자들은 이를 바탕으로 연구 우선순위를 정하는 경우가 많다. 이 현상을 극복한다면 Co-Scientist의 성능을 더욱 향상시킬 수 있다.
- **멀티모달 추론 및 도구 활용 능력의 향상 필요:** 과학 논문에서 가장 흥미로운 데이터는 텍스트가 아닌 그림이나 차트 같은 시각 자료일 수 있다. 최첨단 모델조차도 이러한 데이터를 최적의 방식으로 활용하지 못하는 경우가 많으며, AI Co-Scientist 시스템도 예외는 아니다. 특화된 과학 도구, AI 모델, 데이터베이스와의 통합 및 효과적인 활용 능력을 향상시키기 위한 추가 연구가 필요하다.
- **최신 대형 언어 모델(LLM)의 한계 계승:** LLM이 지닌 사실 오류나 환각의 영향을 받는다. LLM과 웹 검색으로 폭넓은 지식에 접근하지만, 해당 자원의 오류/편향/한계가 전파된다.
- **더 나은 평가 기준과 광범위한 평가의 필요성:** AI Co-Scientist의 평가는 AI 자동 평점, 전문가 리뷰, 특정 실험(in vitro) 검증을 포함하고 있지만, 다양한 생의학 및 과학 분야에 걸친 포괄적이고 체계적인 평가로 시스템의 일반화 가능성을 확인할 수 있다. 또한, 고품질 출판 수준의 엄격함과 세부사항을 충족하도록 개선이 필요하다. 또한, 시스템의 자체 개선에 활용된 Elo 평점보다 더 객관적이고 전문가의 관점을 반영하는 평가 지표가 필요하다.
- **검정(Validation)의 한계:** 현재 AI Co-Scientist는 주로 잠재적인 치료 표적 및 작용 기전을 식별하는 데 초점을 맞추었으며, 약물 전달 시스템의 복잡성은 다루지 않는다. 제약학적 요소는 임상 적용에 매우 중요하지만, 현재 시스템의 범위를 넘어서는 요소이다.

- AI Co-Scientist는 포괄적인 임상시험 설계를 생성하거나, 약물 재창출이나 신약 개발에 적용될 때 약물의 생체이용률, 약물동태학, 복잡한 약물 상호작용과 같은 요소들을 완전히 고려하도록 설계되어 있지 않다. 예측 결과를 실제 임상으로 연결하기 위해서는 전담 중개(translational) 과학 팀이 필요하다. 이러한 한계는 특화된 AI 모델이나 실시간 데이터베이스와 같은 다양한 도구들과 시스템을 지속적으로 개발하고 통합할 것을 요구한다.

6. 안전성과 윤리적 고려사항

AI 시스템이 과학적 발견을 가속화하지만, 과학적 방법론에 대한 영향과는 별개로 안전과 윤리 관련 중요한 과제가 있다. 안전성 측면에서는 이중용도 가능성과 과학적 성과의 악용이 우려되며, 윤리적 측면에서는 특정 과학 분야에서 확립된 윤리 기준을 위반할 수 있다.

- **첨단 AI 기반 과학 연구를 위한 윤리, 정책, 규제의 발전:** 연구 윤리는 과학 활동의 핵심 요소로써, 주된 목표는 사회적으로 긍정적인 영향을 줄 수 있는 연구를 유도하는 것이지만, 동시에 이중용도 가능 지식에 대한 우려도 여전하다. 기본적인 윤리 원칙은 점차적으로 신흥 규제와 조직 내 윤리 검토 과정을 통해 보완되고 있다. 이러한 검토는 연구 제안이 윤리 강령을 따르고 있는지, 현재와 미래의 위험이 무엇인지 평가하는 역할을 한다.

- 자율적 에이전트 AI의 등장으로 인해 과학 및 AI 윤리 정책 역시 변화하고 있다. 이는 변화하는 연구 환경과 AI 에이전트의 위험 요소를 다루기 위해 필수적이다. Co-Scientist처럼 발전된 시스템은 기존의 제한적 AI 모델을 위한 윤리 기준을 넘어서는 새로운 접근이 필요하다. 최근에는 사용자 의도, 도메인, 사회적 영향에 따라 LLM 기반 AI가 과학에 미치는 위험을 구조화하려는 프레임워크도 등장하고 있다.

• **이중용도 위험과 기술적 안전장치:** 설득, 기만, 사이버보안, 자기 확장, 자기 추론 같은 AI agent의 위험한 능력을 평가하는 폭넓은 프레임워크가 개발되고 있으며, 이러한 능력에 대한 폭넓은 평가도 과학 안전성 평가에 통합되어야 한다. 장기적으로는 AI agent가 자체 목표를 가지게 되면서 연구 방향에 영향을 줄 수 있고, 인간이 AI의 조작에 쉽게 노출된다는 점에서 명확한 명령 수행과 가치 정렬 프레임워크가 필수적이다.

- 단기적으로, 비윤리적인 연구 질의, 악의적인 사용자 의도, 그리고 위험하거나 이중용도로 사용될 수 있는 지식을 추출할 가능성에 대응하는 기술적 안전장치가 필요하다. 고도화된 LLM을 '비평(critic)' 또는 '심판(judge)'으로 활용하여 사용자 질의를 평가하는 것과 AI 출력을 감독 메커니즘으로 활용하는 연구가 활발히 진행되고 있다. 하지만 전문적인 분야에서 LLM의 판단이 전문가의 판단과 일치하지 않을 수 있기에 사람의 개입이 필수적이다.

• **과학 AI 시스템의 적대적 공격 대응력:** 적대적 공격에 대한 대비는 기반 모델과 AI 어시스턴트 개발에 있어 매우 중요하다. 지금까지 수작업 진행 레드팀 방식이 여러 취약점을 밝혀냈지만, 최근 greedy 기법, 그래디언트 기반 기법, 진화 알고리즘 등 자동화된 방법으로 프롬프트 접미어를 최적화해 안전장치를 우회하는 기술이 개발되고 있다. 공격은 또한 few-shot 예시, 문맥 내 학습, 멀티모달 입력을 활용하고 있으며, LLM이 다른 LLM에 대한 공격을 생성하고 정교화하는데 여러 단계를 거치는 반복적 형태로 진행될 수 있다. 이러한 공격에 대응하기 위한 방어 기법들도 함께 개발되고 있으며, 이는 특히 agentic AI가 점점 더 보편화될 미래에 더욱 중요하다.

- 기반 모델에 대한 사후 학습이 발전하면서 전반적인 적대적 견고성은 향상될 것으로 기대된다. 하지만, 특정 도메인에서의 악의적 사용을 탐지하고 대응하는 능력은 여전히 과학 분야에 특화된 AI 어시스턴트에 맞는 별도의 개발 및 통합이 필요하다. 또한, 질의 해석, 가설 생성, 내부 사고, 평가, 사용자 질의 처리 등과 같이 반복적 추론을 수행하는 AI 시스템의 경우, 각 구성 요소들은 개별적으로 테스트되어야 한다. 이러한 포괄적인 테스트는 위험한 질의를 어떻게 처리하는지, 가설의 안전성(중간 단계/최종 결과), 그리고 내부 점검 및 필터링 기능의 정확성 등과 같은 모든 잠재적인 실패 가능성을 고려해야 한다.

• **포괄적인 안전 접근 방식의 필요성:** Co-scientist와 같은 과학 AI 어시스턴트는 통합된 설정 가능한 가이드라인이 안전장치에 있어야 한다. 개발자는 커뮤니티 피드백을 빠르게 반영할 수 있도록 유연한 안전 설계에 우선순위를 두어야 한다. 이러한 의미 기반(semantic) 안전장치는 기존의 소프트웨어 안전 조치들과 함께 보완될 수 있다. 예를 들어, 신뢰할 수 있는 테스터 프로그램, 기능의 점진적 출시, 접근 권한 관리, 로그 기록 요청, 불확실한 출력에 대한 수동 검토 플래그 지정 등이 포함된다.

- 이 시스템의 안전성을 보장하기 위해서는 다음과 같이 기존 AI 안전 가이드라인에 부합하는 다방면의 접근 방식이 요구된다: **잠재적인 취약점을 식별하기 위한 포괄적인 위협 모델링/각 위협 요소에 대응하기 위한 방어 메커니즘 구축/광범위한 레드팀 테스트와 보안 점검/취약점 패치 등을 포함한 신속한 대응 절차 마련/지속적인 모니터링과 성능 추적**

- 이 고려사항들은 과학 발전을 위한 책임감 있는 기술 개발, 신중한 운영 및 배포, 적절한 안전장치와 윤리적 가이드라인, 그리고 관련 규정의 철저한 준수가 반드시 필요하다는 점을 강조한다. 또한, 과학 분야에서 AI를 안전하고 윤리적으로 활용하기 위한 모범 사례와 권장 사항을 광범위한 커뮤니티 참여와 포용적인 방식으로 만들어가는 것도 중요하다.

• **Co-Scientist의 안전장치:** 위험 완화를 위해 다음과 같은 안전 메커니즘을 사용하고 있다.

.공개 프론티어 LLM 사용: 폭넓은 안전성 평가와 보호 장치를 갖춘 Gemini 2.0 모델을 사용.

.초기 연구 목표에 대한 안전성 검토: 입력된 연구 목표의 자동 안전성 평가 및 거부 기능.
.연구 가설에 대한 안전성 검토: 연구 목표가 안전하더라도 별도 가설 안전성 검토 진행 및 위험 가설의 토너먼트 제외. 이는 사용자에게도 제공되지 않음.
.연구 방향에 대한 지속적인 모니터링: 메타 리뷰 에이전트의 연구 방향 종합적 검토로 Co-Scientist의 잠재적인 안전 문제 지속 감시 및 위험성 감지 시 사용자에게 경고.
.설명 가능성과 투명성: 모든 시스템 구성 요소의 최종 판단 및 해당 판단의 상세한 추론 과정 제공으로, 판단에 대한 정당성과 감사 실시.
.포괄적인 로그 기록: 모든 시스템 활동이 향후 분석과 감사에 활용될 수 있도록 저장.
.안전성 평가 및 레드팀 테스트: 현재 구현된 위험 연구 목표 탐지 기능의 견고성과 정확성을 검증하기 위한 초기 레드팀 평가 진행되었음(40개 주제 영역/1,200개의 적대적 연구 목표).
.신뢰할 수 있는 테스터 프로그램: Co-Scientist 시스템의 강점과 한계를 보다 엄격히 이해하기 위해, 실제 시스템 사용 연구자들에게 접근 권한을 제공하고, 실사용 기반의 피드백을 수집할 수 있는 '신뢰할 수 있는 테스터 프로그램'이 운영될 예정
- 무엇보다 중요한 점은, Co-Scientist가 항상 인간 전문가의 지속적인 감독 아래 작동하도록 설계되어 있으며, 최종 결정은 언제나 과학자의 전문적 판단에 따라 내려진다.

7. 향후 과제

• **즉각 개선 사항:** Co-Scientist는 초기 개발 단계로써, 문헌 검토 향상, 외부 도구와의 교차 검증, 사실성 확인 개선, 인용 누락 최소화를 통한 관련 연구 인용율 증대 등이 즉각적으로 필요하다. 일관성 검증 도입은 결함 가설에 대한 검토 축소로 시스템 성능 향상이 기대된다.
• **평가 확장:** 자동 문헌 기반 검증과 시뮬레이션 실험을 포함한 보다 객관적인 평가 지표 개발은 주요 과제이다. 기반 LLM이 가지는 편향이나 오류 패턴을 완화하기 위한 방법도 중요하며, 다양한 에이전트 구성 요소 간 상보성과 최적 조합에 대한 분석도 필요하다. 특히 더 많은 주제 전문가들이 참여하고, 고해상도의 다양한 연구 목표를 포함하는 대규모 평가가 중요한 과제이다. 시스템을 다양한 수준(질병 기전부터 단백질 설계까지, 그리고 다른 과학 분야로의 확장)에서 스트레스 테스트하면 추가적인 개선 영역이 드러날 것이다. 또한, 실험실 자원이 제한적인 만큼, 더 나은 평가 프레임워크는 가설 선택을 도와줄 도구가 된다.
• **기능 확장:** 강화학습을 활용하여 가설 순위 매김/제안 생성/진화적 개선에 기여할 수 있다.
- 현재 시스템은 공개 문헌의 텍스트만 분석하지만, 이미지, 데이터 세트, 주요 공공 데이터베이스 등을 통합하면 Co-Scientist의 가설 생성 및 정당화 능력이 크게 향상될 것이다.
- 향후 다단계 실험이나 조건부 논리를 포함한 더 복잡한 실험 설계에 집중하여야한다. Co-Scientist를 실험 자동화와 통합하면 검증을 위한 폐 루프 구축 및 반복적 개선 근거 기반 마련이 가능할 것이다. 또한, 피드백을 제공하고 사용자 연구로부터 인사이트를 얻는 구조화된 사용자 인터페이스 방식은 인간과 AI 간 협업 효율성을 높여줄 것이다.

8. 논의

Co-Scientist는 과학적 발견을 가속화하려는 초기 시도 에이전트형 AI 시스템으로써, 특화된 에이전트들을 활용하여 '생성-토론-진화' 방식으로 가설 생성 및 반복적 정제가 이루어진다. 이는 연구 가설 생성을 위한 자기 개선 순환 구조를 만들고, 자동 평가 지표를 기준으로 성능이 측정된다. 또한 추론 시 테스트 시점 연산 확장의 이점을 보여준다.
무작위 가설 생성 방식 대신, 이 시스템은 과학적 추론의 핵심 요소들을 지능적으로 모방하도

록 설계되어 있다. 3장에서와 같이, Co-Scientist는 과학적 토론, 토너먼트, 반복적 정제, 인간의 피드백 루프 등 원칙에 기반한 내부 메커니즘을 통해 가설들의 품질을 점진적으로 향상시키고, 고품질의 정제된 가설로 수렴하도록 한다.

- **AI Co-Scientist의 참신함-기존 증거 기반:** Co-Scientist는 방대한 문헌을 종합하고 잠재적인 연관성을 식별함으로써 새로운 과학적 가설을 생성하고 새로운 통찰을 발견하도록 한다. 현재 주요 활용은 점진적인 발전을 가능하게 하는 데 있지만, 향후에는 탐색적이거나 획기적인 연구를 지원할 수도 있다. 연구자가 복잡하고 창의적인 사고를 요구하는 열린 연구 목표를 정의하는 경우, 시스템은 다양한 신뢰 수준의 결과물을 생성할 수 있으므로 해당 결과물에 대한 엄격한 검증과 분야 전문가의 비판적 평가가 필수적이다. 이 시스템은 인간의 과학적 사고를 대체하기 위한 것이 아니라 보완하기 위한 것으로, 연구자가 생성된 통찰에 대한 지적 통제력을 유지하면서 발견을 가속화하도록 설계되었다.
- **Co-Scientist 제안 가설에 대한 다중 실험적 검증:** Co-Scientist가 생성한 가설이 여러 실험실에서 검증되었음을 보여주었다. 그 가설들은 AML 치료 약물 재창출, 간 섬유증의 새로운 후생유전학적 치료 표적 제안, 항균제 내성 분야의 박테리아 종 사이에서 cf-PICI 전달 메커니즘에 대한 새로운(아직 발표되지 않은) 발견이다. 이처럼 다양한 과학적 복잡성을 가진 초기 결과는 Co-Scientist가 지닌 다양한 생의학 분야에서의 잠재력을 보여준다.
- **과학적 추론 선지식과 귀납적 편향을 활용한 테스트 시점 연산 확장:** Co-Scientist는 특별한 사전 학습, 사후 학습, 또는 강화학습 프레임워크의 요구 없이 기본 LLM의 기능을 그대로 활용하였다. 즉, 해당 모델들이 업데이트되면 Co-Scientist는 재학습이 없어도 계산 효율성과 일반화 가능성 면에서 이점을 누릴 수 있다. 이 시스템의 아키텍처는 자기 대전, 내부 일관성 검사, 토너먼트 기반 순위 결정 구조를 포함하고 있어 가설 생성, 평가, 개선을 반복할 수 있다. 이러한 자기 진화 기능은 데이터베이스 질의 등 외부 도구의 통합으로 더욱 향상될 수 있으며, 이를 통해 Co-Scientist는 기존 지식에 기반하여 자신의 제안을 구체화하고 새로운 연결고리를 찾아낼 수 있다. 앞으로는 Co-Scientist가 자체적으로 생성한 데이터와 토너먼트 순위를 활용하여 강화학습으로 시스템 전체를 개선할 가능성도 열려 있다.
- **최첨단 LLM의 발전과 AI Co-Scientist:** Co-Scientist 시스템에 활용된 LLM들은 추론, 논리적 사고, 과학 문헌 이해 등의 측면에서 지속적으로 빠르게 성능이 향상되고 있다. 즉, 특정 모델에 종속되지 않는 구조이기에, LLM의 성능이 향상될수록 도구의 최적 활용을 포함한 새로운 연구 방향도 가능하게 될 것으로 예상된다.
- **약물 재창출 및 신약 개발에 대한 시사점:** Co-Scientist를 약물 후보 선정 과정에 통합하여 근거 기반 약물 재창출에 중요한 진전이 이루어진 것처럼, Co-Scientist는 다양한 생물의학 및 과학 분야에 중대한 진전을 가져왔다. 단순한 문헌 검색을 넘어서, Co-Scientist는 분자 경로, 기존 전임상 증거, 및 잠재적인 치료 적용을 연결하여 새로운 기전적 통찰을 종합하고, 이를 구조화된 검증 가능한 구체적 목표로 만든다. 이러한 능력은 문헌에 기반한 논리를 제공하고, 검증을 위한 구체적인 실험적 접근을 제안한다는 점에서 연구자에게 아주 유용하다. 특히, Co-Scientist가 생성한 구조화된 결과물은 자비 치료를 위한 단일 환자용 임상시험 신약 신청서를 개발하는 데 활용될 수 있다. 기전적 증거, 관련 전임상 데이터, 제안된 모니터링 항목들을 체계적으로 제시하여, Co-Scientist는 기존 치료법이 더 이상 효과를 보이지 않고 임상시험에도 참여할 수 없는 불응성 질환 환자를 위한 합리적인 치료 프로토콜을 개발하게 한다. 이는 특히 희귀하거나 공격적인 질환처럼 전통적인 신약 개발 일정이 환자의 긴급한 필요를 따라가지 못하는 경우에 중요하다. 이 플랫폼은 안전성 고려사항과 모니터링 지표

를 포함한 근거 기반 치료 가설을 신속하게 생성하기 때문에, 자비 치료 신청에 있어 임상의와 규제 당국이 과학적 엄밀성을 유지한 채 정보에 기반한 결정을 내릴 수 있도록 돕는다.

- Co-Scientist를 약물 재창출에 적용하는 것은 특히 희귀질환 적응증에 대한 풍부한 안전성과 임상 데이터를 보유하고 있는 희귀질환 치료제에게 매우 매력적인 기회를 제공한다. 3상 임상시험은 수억 달러가 들므로, 이미 잘 특성화된 치료제를 재활용하는 것은 다양한 질환에 대해 치료 옵션을 확장하는 효율적인 방안이다. 이는 희귀질환 치료제가 종종 다른 질환에서도 작용할 근본적인 생물학적 경로를 표적으로 하지만 연관성은 기존의 전통적인 연구 방식으로는 즉각적으로 드러나지 않기 때문이다. Co-Scientist는 기존의 임상 데이터, 안전성 결과, 기전적 통찰을 체계적으로 평가하여, 이미 진행된 약물 개발 및 안전성 검증의 투자를 활용하면서 새로운 유망 치료 적용을 찾아낼 수 있다. 이는 기존 치료제의 활용도를 극대화하는 동시에, 더 넓은 환자 집단의 의료적 수요를 더 빠르게 해결하는 데 기여할 수 있다.
- 더 나아가, Co-Scientist는 간 섬유증에 대한 타깃 발굴(target discovery) 초기 연구에서 입증된 바와 같이, 신약 개발의 전 범위에 걸쳐 잠재적인 영향력을 가진다.

• **자동화 편향과 과학적 창의성의 영향:** 생물의학 및 과학 분야에서 AI의 잠재력 실현을 위해서는, 잠재적 문제점이 사전에 해결되어야 한다. 협업 AI 시스템에서 AI 생성제안에 과도하게 의존하면 비판적 사고가 약화되고 연구는 동질화 된다. AI의 창의성과 아이디어 발상에 대한 영향의 연구 결과들은 혼재되어 있다. 일부는 아이디어의 동질화를 시사하며, 다른 연구들은 명확한 결론을 내리지 못하고 있다. 또한, 유사한 데이터를 학습한 LLM의 성공/실패 상관관계는 과학적 탐구의 범위를 좁게 만든다. 더불어, AI 시스템의 사각지대와 연구 분야에 따른 성능 편차 역시 고려해야 하기에, 확장 가능한 사실 검증 및 검토 방법과 함께 동료 평가와 잠재적 편향에 대한 신중한 고려가 필수적이다. Co-Scientist 같은 시스템은 이러한 위험을 완화할 수 있도록 신중하게 설계되고 사용되어야 한다.

• **과학적 발견과 형평성 증진 촉매로서의 AI:** 이러한 위험에도 불구하고, AI는 과학 정보접근 민주화 및 발견 가속화의 엄청난 잠재력을 지녀, 역사적 소외 혹은 자원 부족 지역에 큰 혜택을 줄 것이다. AI는 과학 발전 "조수"를 끌어올려, 뒤처졌던 이들도 함께 띄워줄 수 있다. 이를 실현하려면, 오류 가능성을 최소화하면서 아이디어 창출과 혁신을 촉진하도록 AI 시스템에 대한 전략적인 투자와 세심한 조정이 필요하다. 역사적으로 소외되었던 연구 주제에 집중하고, 사전 데이터 양이 상이한 다양한 과학 분야 전반에서의 성능 차이를 해소하여야 한다. 현재는 점진적인 아이디어나 연구 가설 생성 정도이지만, 앞으로는 독창적이고, 비정통적이며, 혁신적인 과학 이론을 생성하고자한다. 이 도전과제를 선제적으로 해결한다면, AI는 모든 과학자에게 강력한 도구가 되어 공정하고 혁신적인 과학 탐구의 미래를 열 것이다.

9 결론

AI Co-Scientist는 과학자들의 역량을 보완하고 과학적 발견을 가속화하는 데 있어 유망한 진전을 가져왔다. 다양한 과학 및 생물의학 분야에서 새로운 검증 가능한 가설을 생성하는 기능과-그 중 일부는 실험적 증거로 뒷받침됨- 연산에 의해 스스로를 반복적으로 개선할 수 있는 기능은, 인간 건강, 의학, 과학 분야의 중대한 과제를 해결하려는 과학자들의 노력을 실질적으로 가속화할 수 있음을 보여준다. 이러한 혁신은 수많은 질문과 기회를 가져왔다. 과학이 지닌 실증적이고 책임 있는 접근 방식을 AI Co-Scientist 시스템 자체에 적용함으로써, 인간 중심의 협업 AI 시스템이 인간의 창의력을 어떻게 증진시키고 과학적 발견을 가속할 수 있는지를 포함하여 AI의 잠재력을 안전하게 탐구할 수 있을 것이다.

Google DeepMind의 AlphaEvlove와 AlphaGenome

취지: Google DeepMind는 이세돌과의 바둑 대결에서 승리를 거둔 AlphaGo 이후에도 꾸준히 혁신적인 연구 결과들을 발표하고 있다. 2025년에 발표한 대표적인 사례로서, 6월에 발표한 AlphaEvolve와 7월에 발표한 AlphaGenome에 대하여 조사하였다.

AlphaEvlove는 LLM기반 진화형 AI 코딩 에이전트이며, 대규모 언어모델(LLM)과 진화 알고리즘을 활용하여 복잡한 알고리즘을 설계하고 최적화 해준다. AlphaEvolve로 Google의 데이터센터 스케줄링과 TPU 설계를 최적화하고 LLM의 학습속도를 가속화 하였다.

AlphaGenome은 DNA 서열의 유전자 기능 예측과 질병 생물학 연구에 혁신을 가져왔다. 즉, 최대 100만 염기쌍의 DNA 서열을 입력받아 염기 수준의 고해상도 생명현상을 예측하게 해주며, 다양한 유전자 조절 변이의 효과를 예측한다.

조사자료:

A. Novikov et al., “AlphaEvolve: A coding agent for scientific and algorithmic discovery,” https://doi.org/10.48550/arXiv.2506.13131

Žiga Avsec et al., “AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model” Preprint, Google Deepmind

Ewen Callaway, “DeepMInd’s new AI tackles the ‘Dark Matter’ in Our DNA,” Nature, vol. 643, pp. 17~18, 3 July 2025.

https://deepmind.google/discover/blog/alphagenome-ai-for-better-understanding-the-genome/

“AlphaEvolve:
A Coding Agent for Scientific and Algorithmic Discovery”

요약: AlphaEvolve는 진화적 코딩 에이전트로서, 기존 최첨단 대규모 언어 모델(LLM)의 능력을 크게 향상시켜, 개방형 과학 문제 해결이나 핵심 계산 인프라 최적화와 같은 매우 도전적인 작업을 수행할 수 있도록 한다. AlphaEvolve는 LLM들로 구성된 자율 파이프라인을 조율하여 알고리즘을 직접적으로 수정함으로써 개선하는 역할을 수행한다. 이 시스템은 진화적 접근 방식을 사용하며, 하나 이상의 평가자로부터 지속적으로 피드백을 받아 알고리즘을 반복적으로 향상시키며, 궁극적으로 새로운 과학적·실용적 발견으로 이어질 수 있다. 여기서는 이 접근 방식이 다양한 중요 계산 문제에 널리 적용 가능하다는 점을 입증한다. Google의 대규모 계산 스택에서 핵심 구성 요소를 최적화할 때, AlphaEvolve는 데이터 센터를 위한 보다 효율적인 스케줄링 알고리즘을 개발했다. 하드웨어 가속기의 회로 설계에서 기능적으로 동일하지만 더 간단한 방식을 찾아냈고, AlphaEvolve 자체를 구동하는 LLM의 학습 속도 또한 가속화했다. 더군다나, AlphaEvolve는 수학 및 컴퓨터 과학 분야의 문제에서 기존 최첨단 솔루션을 능가하는 새롭고 수학적으로 입증된 올바른 알고리즘을 발견함으로써, 기존 자동 발견 방법(Romera-Paredes et al., 2023)의 범위를 크게 확장했다. 특히 주목할 성과로, AlphaEvolve가 4×4 복소수 행렬 곱셈을 48회의 스칼라 곱셈으로 수행하는 방법을 찾아내었다. 이는 이 분야에서 Strassen 알고리즘 이후 56년 만에 이루어진 개선이다. AlphaEvolve 및 이와 같은 코딩 에이전트가 과학 및 계산의 다양한 분야에서 문제 해결 능력을 크게 향상시킬 것이다.

1 서론

가치가 높은 새로운 지식을 발견하는 일-예를 들어, 혁신적인 과학적 발견을 하거나 상업적으로 가치 있는 알고리즘을 개발하는 것-은 일반적으로 아이디어 구상, 탐색, 가능성 없는 가설에 대한 역추적, 실험, 검증 등의 오랜 과정이 필요하다. 최근 이 과정을 부분적으로 자동화하기 위해 대규모 언어 모델(LLM)을 활용하려는 시도가 많아지고 있다. 이러한 시도에 대한 기대는, 최신 LLM들이 테스트 시점의 연산 자원을 활용해 능력을 확장하여 보여주는 놀라운 성능, 그리고 언어 생성과 행동을 결합한 에이전트의 등장 등에 의해 뒷받침되고 있다. 이러한 발전은 여러 기존 벤치마크에서 성능을 향상시켰을 뿐 아니라, 가설 생성이나 실험 설계처럼 발견 중심의 작업을 가속화하기도 했다. 하지만 LLM 기반 파이프라인이 완전히 새로운 과학적 또는 실용적 발견에까지 도달하게 하는 것은 여전히 도전적인 과제로 남아 있다.

여기서는 진화 연산(evolutionary computation)과 LLM 기반 코드 생성을 결합한 LLM 기반 코드 초최적화(superoptimization) 에이전트 “AlphaEvolve”를 알아본다. AlphaEvolve는 발견 후보군의 자동 평가가 가능한 과학 및 공학 분야의 폭넓은 문제를 다루는 데 초점을 맞추었다. 이 시스템은 새로운 수학적 객체나 실용적인 휴리스틱과 같은 발견 후보를 알고리즘의 형태로 표현하고, 여러 개의 LLM을 활용하여 이러한 알고리즘들을 생성, 평가, 진화시킨다. LLM이 주도하는 이 진화 과정은 코드 실행 및 자동 평가를 통해 실질적으로 뒷받침된다. 이러한 평가 메커니즘은 AlphaEvolve가 기반 LLM의 부정확한 제안을 피할 수 있게 해준다.

AlphaEvolve의 진화적 과정은 최신 LLM의 피드백 반응 능력을 활용하여, 초기 후보군과 문법적으로나 기능적으로 상당히 다른 새로운 후보를 발견한다. 이 방식은 새로운 알고리즘을

발견하는 것이 본질적인 목표인 문제뿐만 아니라, 해결하고자 하는 대상이 알고리즘 그 자체는 아니지만, 그 해법을 구성하거나 찾는 과정을 알고리즘으로 기술할 수 있는 다양한 문제들에도 적용 가능하다. 후자의 경우, 알고리즘을 찾는 것은 수단적인 목표일 뿐이지만, 해답 자체를 직접 탐색하는 것보다 놀라울 만큼 효과적인 전략이다.

진화적 방법과 코딩 LLM을 결합하는 아이디어는 이전에도 여러 특화된 환경에서 연구되었다. 특히, AlphaEvolve는 수학적 객체를 구성하거나 온라인 알고리즘을 구동하는 휴리스틱을 발견하는 데 LLM이 주도하는 진화를 사용한 FunSearch의 실질적인 확장판으로 볼 수 있다 (Table 1). 이와 유사한 접근 방식은 시뮬레이션 로봇을 위한 정책 발견, 기호 회귀(symbolic regression), 조합 최적화를 위한 휴리스틱 함수 합성과 같은 작업에도 활용되었다. 이러한 시스템들과 달리, AlphaEvolve는 최신(state-of-the-art, SOTA) LLM을 활용하여 여러 함수와 구성 요소에 걸친 복잡한 알고리즘을 구현하는 대규모 코드 조각을 진화시킨다. 그 결과, AlphaEvolve는 규모와 일반성 면에서 기존 시스템들을 크게 능가하는 성능을 보여준다.

Table 1 | Capabilities and typical behaviours of *AlphaEvolve* and our previous agent.

FunSearch [83]	*AlphaEvolve*
evolves single function	evolves entire code file
evolves up to 10-20 lines of code	evolves up to hundreds of lines of code
evolves code in Python	evolves any language
needs fast evaluation (≤ 20min on 1 CPU)	can evaluate for hours, in parallel, on accelerators
millions of LLM samples used	thousands of LLM samples suffice
small LLMs used; no benefit from larger	benefits from SOTA LLMs
minimal context (only previous solutions)	rich context and feedback in prompts
optimizes single metric	can simultaneously optimize multiple metrics

자동 평가 지표의 사용은 AlphaEvolve에 중요한 이점을 제공하지만, 동시에 하나의 한계점이기도 하다. 특히, 수동 실험이 필요한 작업들은 AlphaEvolve의 적용 범위 밖이다. 그러나 수학, 컴퓨터 과학, 시스템 최적화 분야의 문제들은 일반적으로 자동화된 평가 지표를 활용할 수 있기 때문에, AlphaEvolve는 이러한 분야에 집중하여 개발되었다. 구체적으로, AlphaEvolve는 알고리즘 설계와 구성적 수학(constructive mathematics)의 잘 알려진 여러 미해결 문제 해결과 Google의 대규모 계산 스택에서 핵심 계층의 최적화에 진전을 이루었다.

알고리즘 설계 분야에서, 행렬 곱셈을 빠르게 수행하는 알고리즘을 발견하는 근본적인 문제를 다루었다. 이 문제는 과거에 보다 특화된 AI 접근법이 적용된 바 있다. AlphaEvolve는 범용적인 시스템임에도 불구하고 14개의 행렬 곱셈 알고리즘에서 SOTA를 개선하는 성과를 보였다. 특히 4×4 행렬의 경우, AlphaEvolve는 4×4 복소수 행렬 곱셈을 48회의 곱셈만으로 수행하는 알고리즘을 찾아내어, Strassen의 1969년 알고리즘을 개선하는 데 성공했다.

수학에서는, 주어진 수학적 정의에 따라 기존에 알려진 모든 구성보다 더 나은 특성을 가진 구성을 발견함으로써 진전을 이룰 수 있는 다양한 미해결 문제들을 다루었다. 이러한 문제들 중 50개 이상의 문제에 AlphaEvolve를 적용하였다. 약 75%의 문제에서는 현재 알려진 최상의 구성과 동등한 결과를 얻었으며, 약 20%의 문제에서는 SOTA를 능가하는 새로운 구성을 발견하였다. 이는 수학적으로도 향상이 입증된 결과이다. 이에는 Erdős가 제안한 최소 중첩 문제에 대한 개선과, 11차원 Kissing Numbers 문제에 대한 향상된 구성도 포함된다.

마지막으로 Google의 컴퓨팅 스택의 다양한 계층에 걸친 다음 네 가지 공학적 문제에 AlphaEvolve를 적용했다: “Google 클러스터 관리 시스템을 위한 스케줄링 휴리스틱 발견”, “LLM 학습에 사용되는 행렬 곱셈 커널 최적화”, “TPU 내부의 산술 회로 최적화”, “트랜스포머의 어텐션 실행 시간 최적화”. 이러한 구성 요소들은 장기간에 걸쳐 반복적으로 실행되므로, 작은 개선이라도 매우 높은 가치를 지닌다.

2 AlphaEvolve

AlphaEvolve는 LLM에 대한 질의 등을 포함한 일련의 계산 과정을 자율적으로 조율하는 파이프라인을 구성하여, 사용자가 지정한 작업을 해결하는 알고리즘을 생성하는 코딩 에이전트이다. 상위 수준에서 이 조율 절차는 주어진 작업에 연결된 자동 평가 지표의 점수를 향상시키는 프로그램들을 점진적으로 개발해내는 진화 알고리즘이다. 그림1: 개요, 그림2: 확장 구조

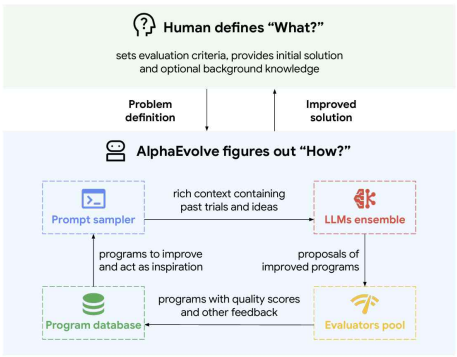


Figure 1 | *AlphaEvolve* high-level overview.

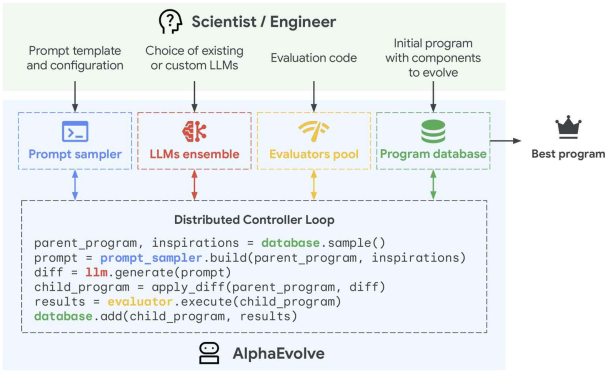


Figure 2 | Expanded view of the *AlphaEvolve* discovery process. The user provides an initial

2.1 작업 명세(Task Specification)

평가(Evaluation): AlphaEvolve는 기계적 채점이 가능한 문제를 다루기 때문에, 사용자는 생성된 해법을 자동으로 평가할 수 있는 메커니즘을 제공해야 한다. 이 메커니즘은 해법을 스칼라 평가 지표들의 집합으로 매핑하는 함수 h로 주어지며, 평가지표를 최대화하는 해법을 구하여야 한다. h는 고정된 입력/출력 형식을 갖는 함수 evaluate로 구현되며, 스칼라 값들의 dictionary를 반환한다. 응용 분야에 따라 이 함수의 실행은 단일 장치에서 몇 초만에 끝날 수도 있고, 광범위한 계산을 유발할 수도 있다. 수학적 문제의 경우, 함수 h는 일반적으로 매우 단순하다. 예를 들어, 주어진 성질을 만족하는 가능한 한 큰 그래프를 찾고자 할 때, h는 진화된 코드를 호출해 그래프를 생성하고, 해당 성질을 만족하는지 확인한 뒤, 단순히 그래프의 크기를 점수로 반환한다. 더 복잡한 경우에는, h 함수가 진화된 탐색 알고리즘을 실행하거나, 머신러닝 모델을 학습시키고 평가하는 작업을 포함할 수 있다.

API: 코드베이스 전반에 걸쳐 여러 구성 요소를 동시에 진화시키는 작업을 지원하기 위해, AlphaEvolve는 코드 블록에 시스템이 진화시킬 대상으로 지정할 수 있는 입력 API를 제공한다. 이러한 설계는 기존 코드베이스와의 통합을 용이하게 하며, 코드에 주석 형태로 특수 마커(# EVOLVE-BLOCK-START 및 # EVOLVE-BLOCK-END)만 추가하면 된다. 이러한 진화 블록 안에 포함된 사용자 제공 코드는 AlphaEvolve가 개선할 초기 해법으로 사용되며, 그 외의 코드는 진화된 코드 조각들을 연결하는 skeleton 역할을 하여, 이를 evaluate 함수에서 호출할 수 있도록 한다. 이 초기 구현은 완전한 형태여야 하지만, 매우 단순한 수준이어도 무방하다. 예를 들어, 적절한 타입의 상수를 반환하는 한 줄짜리 함수들로 구성될 수도 있다.

추상화 수준의 선택에 대한 유연성: AlphaEvolve는 같은 문제에 대해서도 매우 다양한 방식으로 적용될 수 있다-특히, 진화된 프로그램이 최종 출력이 아니라 해결책을 발견하기 위한 수단일 경우-. 예를 들어, AlphaEvolve는 다음과 같은 다양한 방식으로 해결책을 진화시킬 수 있다: 고전적인 진화 알고리즘처럼 문자열 형태의 해법 표현을 직접 진화; 해결책을 처음부터 구성하는 방식을 명시한 특정 형태의 함수를 진화; 제한된 계산 자원 내에서 해결책을 찾기 위한 맞춤형 탐색 알고리즘을 진화; 중간 해법과 탐색 알고리즘을 함께 공진화(co-evolve); 각 탐색 알고리즘이 특정 중간 해법을 더욱 개선하는 데 특화되도록 한다. 다양한 문제에 따라 서로 다른 추상화 수준이 더 효과적이라는 것을 발견했다. 예를 들어, 해결책이 고도의 대칭성을 갖는 문제의 경우, 보다 간결한 형태로 표현되는 경향이 있는 생성자 함수를 진화시키는 것이 유리하다고 가정한다. 반면, 비대칭적인 해결책을 가지는 문제의 경우에는 맞춤형 탐색 알고리즘을 진화시키는 방식이 더 효과적이다.

2.2 프롬프트 샘플링(Prompt Sampling)

AlphaEvolve는 SOTA LLM을 활용하기 때문에, 다양한 형태의 맞춤화와 긴 콘텍스트 제공을 주요 진화 프롬프트의 일부로 지원한다. 이 프롬프트는 프로그램 데이터베이스에서 샘플링된 이전에 발견된 여러 해법들과 특정 해법에 대해 어떻게 변경을 제안할 것인지에 대한 시스템 지침 등으로 구성된다. 이러한 핵심 요소 외에도, 사용자는 프롬프트를 다음과 같은 방식으로 자신의 필요에 맞게 더욱 세밀하게 조정할 수 있다:

·**명시적 콘텍스트:** 해결하려는 문제에 대한 세부 정보로, 사람이 작성한 고정된 지침, 수식, 코드 조각, 또는 관련 문헌(예: PDF 파일) 등이 포함될 수 있다.

·**확률적 형식화:** 다양성을 높이기 위해 사람이 제공한 여러 대안들을 갖는 템플릿 자리 표시자로, 이는 별도의 설정 파일에 정의된 확률 분포에 따라 구체적으로 생성(instantiated)된다.

·**렌더링된 평가 결과:** 일반적으로 프로그램, 실행 결과, 그리고 evaluate 함수에 의해 할당된 점수를 지닌다.

·**메타 프롬프트 진화:** 추가 프롬프트 생성 단계에서 LLM이 직접 제안한 지침 및 콘텍스트를 포함하며, 이는 해결책 프로그램과 유사한 방식으로 별도의 데이터베이스에서 공동 진화된다.

2.3 창의적 생성 (Creative Generation)

진화 과정을 이끌기 위해 AlphaEvolve는 SOTA LLM의 기능을 활용한다. 이들 LLM의 주요 역할은 이전에 개발된 해법들에 대한 정보를 이해하고, 이를 개선하기 위한 새롭고 다양한 방법들을 제안하는 것이다. AlphaEvolve는 모델에 구애받지 않는(model-agnostic) 구조를 가지고 있지만, 실험(ablation) 결과에 따르면, 기반이 되는 LLM의 성능이 향상될수록 AlphaEvolve의 성능도 함께 향상되는 것을 확인할 수 있다(4절 참조).

출력 형식(Output format): LLM에게 기존 코드를 수정하도록 요청할 때, 특히 대규모 코드베이스 내에서는, 수정 사항을 특정 형식의 diff 블록 시퀀스로 제공하도록 요청한다:

```
<<<<<<< SEARCH
  # Original code block to be found and replaced
=======
  # New code block to replace the original
>>>>>>> REPLACE
```

여기서 <<<<<<< SEARCH와 ======= 사이의 코드는 현재 프로그램 버전에서 정확히 일치해야 하는 코드 구간이며, =======와 >>>>>>> REPLACE 사이의 코드는 기존 코드를 대체할

새로운 코드 구간이다. 이 형식은 코드의 특정 부분만을 정확하게 업데이트할 수 있게 해준다. 진화 중인 코드가 매우 짧거나, 작은 수정보다 전체 재작성이 더 적절한 경우에는, AlphaEvolve를 설정하여 diff 형식 대신 전체 코드 블록을 직접 출력하도록 지시할 수 있다.

사용된 모델(Models used): AlphaEvolve는 여러 개의 LLM으로 구성된 앙상블을 사용한다. 구체적으로는, Gemini 2.0 Flash와 Gemini 2.0 Pro의 조합을 활용하는데, 이러한 앙상블 접근 방식은 계산 효율성과 생성된 해법의 품질 간의 균형을 가능하게 한다. Gemini 2.0 Flash는 지연 시간이 짧아 더 빠른 속도로 후보를 생성할 수 있으므로, 단위 시간당 탐색되는 아이디어의 수를 증가시킨다. 보다 높은 성능을 가진 Gemini 2.0 Pro는 간헐적으로 고품질의 제안을 생성하여, 진화 탐색을 크게 진전시키거나 중요한 돌파구를 마련할 수 있다. 이러한 전략적 조합은, 많은 수의 아이디어를 평가하면서도, 보다 강력한 모델을 통해 실질적인 개선 가능성을 유지함으로써, 전체적인 발견 과정을 최적화한다.

2.4 평가(Evalution)

AlphaEvolve의 진화 과정을 추적하고, 다음 세대로 전달할 아이디어를 선택하기 위해, LLM이 제안한 각 새로운 해법은 자동으로 평가된다. 원칙적으로 이 과정은 생성된 해법에 대해 사용자가 제공한 평가 함수 h를 실행하는 것에 해당하지만, 실제로는 이 평가 과정을 더 유연하고 효율적으로 만들기 위해 선택적으로 사용할 수 있는 다양한 메커니즘을 지원한다:

·**평가 캐스케이드(Evaluation cascade, 가설 검정):** 사용자는 난이도가 점점 높아지는 테스트 사례 ensemble을 지정할 수 있으며, 새로운 해법이 이전 단계들에서 충분히 유망한 결과를 낼 경우에만 다음 단계로 평가가 진행된다. 이는 가능성이 낮은 해법들을 신속히 걸러내도록 해준다. 또한 새로운 해법은 소규모 테스트를 통해 초기 평가를 받으며, 문제가 있는 프로그램을 조기에 걸러내기 위한 사전 필터링 단계로 활용된다.

·**LLM 생성 피드백:** 일부 응용에서는 사용자가 제공한 평가 함수 h로는 정확하게 반영하기 어려운-예를 들어 발견된 프로그램의 단순성-특정 특성이 존재한다. 이러한 특성들은 별도의 LLM 호출을 통해 평가되어 점수 딕셔너리에 추가될 수 있으며, 이를 통해 진화를 유도하거나, 기준을 충족하지 않는 해법을 배제하는 데 사용될 수 있다.

·**병렬 평가:** AlphaEvolve의 샘플 효율성 덕분에, 새로운 해법 하나를 평가하는 데 약 100 compute-hours를 투입하는 것이 가능하지만, 개별 평가가 병렬 처리되지 않아 실제 경과 시간이 줄어들지 않으면, 새로운 세대가 등장하는 속도가 느려져 진화 알고리즘이 연속적인 변이를 여러 차례 적용하는 능력이 제한된다. 많은 응용 분야에서 평가는 매우 쉽게 병렬화할 수 있는데, 이러한 작업을 평가 클러스터에 대한 비동기 호출을 통해 분산 처리할 수 있다.

다중 점수(Multiple scores): AlphaEvolve는 사용자가 제공한 여러 평가 지표에 대해 동시에 최적화할 수 있도록 지원한다. 이 기능은 본질적인 가치뿐만 아니라 수단적인 가치도 지닌다. 여러 응용 분야에서 우리는 실제로 복수의 평가 지표에 대해 강한 해법을 개발하는 것에 관심이 있지만, 한 가지 지표만 특별히 중요하더라도 여러 지표를 함께 최적화하는 것이 단일 목표 지표에 대한 결과를 개선하는 데 도움이 되는 경우가 많다. 이는 아마도, 서로 다른 평가 기준에서 우수한 프로그램들이 각기 다른 구조나 논리를 가지기 때문일 것이다. 그리고 이러한 다양한 고성능 프로그램들—각각이 '좋음'의 다른 정의를 대표하는—의 사례를 LLM에 제공하는 프롬프트에 포함시킴으로써, 더 다양하고 폭넓은 후보 해법 생성을 자극할 수 있다. 이로 인해 목표 지표에 대해 매우 효과적인 새로운 접근법을 발견할 가능성이 높아진다.

2.5 진화(Evolution)

진화 과정 동안 AlphaEvolve는 평가 결과(점수와 프로그램 출력)가 첨부된 해법들을 지속적으로 생성한다. 이 해법들은 진화 데이터베이스에 저장되어, 이전에 탐색한 아이디어를 최적으로 재활용하게 한다. 이 데이터베이스 설계에서 핵심 과제는 최고의 프로그램을 지속적으로 개선하기 위한 exploitation과 전체 탐색 공간을 탐험하기 위한 exploration를 균형 있게 유지하는 것이다. AlphaEvolve에서는 진화 데이터베이스가 MAP 엘리트 알고리즘과 섬 기반 개체군 모델을 결합한 방식에서 착안한 알고리즘을 구현하고 있다.

2.6 분산 파이프라인 (Distributed pipeline)

AlphaEvolve는 비동기적 계산 파이프라인으로 구현되었으며, Python의 asyncio 라이브러리를 사용한다. 이 파이프라인에서는 많은 계산들이 동시에 실행되며, 각 계산은 다음 단계가 아직 완료되지 않은 다른 계산의 결과에 의존할 경우 해당 결과가 나올 때까지 대기한다. 구체적으로, 이 비동기 파이프라인은 컨트롤러, LLM 샘플러, 그리고 평가 노드로 구성되어 있다. 전체 파이프라인은 개별 계산의 속도보다는 처리량을 최적화하는 데 중점을 두어, 주어진 전체 계산 예산 내에서 최대한 많은 아이디어를 제안하고 평가할 수 있도록 설계되었다.

3 결과

3.1 텐서 분해를 통한 새로운 알고리즘 발견으로 더 빠른 행렬 곱 구현

머신러닝 계산 가속화에서부터 현실적인 컴퓨터 그래픽 구현에 이르기까지, 행렬 곱하기는 컴퓨터 과학 전반의 수많은 핵심 알고리즘과 응용을 뒷받침하는 기본 연산이다. 두 행렬의 곱셈을 수행하는 다양한 알고리즘들이 주어진 3차원 텐서를 랭크 1 텐서로 분해하는 방식으로 표현될 수 있다. 이러한 분해의 랭크(항의 개수)는 행렬 곱을 계산하는 데 필요한 스칼라 곱셈의 수를 정확히 나타낸다. 따라서 더 빠른 행렬 곱 알고리즘을 개발하려면, 특정 텐서의 저랭크 분해를 찾아야 한다. 이 문제는 특화된 교대 최소자승법, 딥 강화학습, 맞춤형 탐색 알고리즘 등 다양한 접근 방식으로 오랫동안 연구되어 왔다. 그럼에도 불구하고, 단순한 3×3 행렬 곱셈의 경우 달성 가능한 최소 랭크조차 아직 알려져 있지 않을 정도로, 이 문제는 여전히 매우 어렵고 난해한 과제로 남아 있다.

문제 설명과 함께 초기화 함수, 재구성 손실 함수, Adam 옵티마이저를 포함한 표준 경사 기반 알고리즘에서 출발하여, AlphaEvolve는 기존 방법들을 능가하는 정교한 텐서 분해 알고리즘을 스스로 개발할 수 있다. 각 진화된 프로그램을 평가하기 위해, 여러 개의 행렬 곱셈 목표를 선택하고, 섹션 2.4에서 설명한 평가 캐스케이드를 활용해 여러 랜덤 시드로 초기화된 상태에서 알고리즘을 실행한다. 성능 평가는 다음 두 가지 지표로 이루어진다: 각 목표에 대해 달성된 최소 랭크(최소 스칼라 곱셈 수), 해당 랭크를 달성한 시드 비율(성공률). 이러한 결과는 AlphaEvolve가 hill-climbing 전략을 사용할 수 있도록 신호로 작용한다. 또한 분해의 정확성을 보장하고 수치 오류 가능성을 피하기 위해, 평가 시 각 요소를 가장 가까운 정수 혹은 0.5 단위의 수로 반올림한다. 그리고 알고리즘이 정수 또는 거의 정수에 가까운 해법을 생성하도록 유도하기 위해, 이러한 요구 사항을 LLM 프롬프트에 자연어로 포함시킨다.

AlphaEvolve가 개발한 알고리즘들은 14개의 서로 다른 행렬 곱셈 목표에 대해 성능을 향상시켰다. 특히, 두 개의 4×4 행렬 곱하기의 경우, Strassen의 알고리즘을 재귀적으로 적용하면 랭크가 49인 알고리즘이 나오며, 이 알고리즘은 모든 field에서 작동한다. 반면, 2개의 원소를

가지는 F_2에서의 특정한 경우에 대해서는, Fawzi가 랭크 47의 알고리즘을 발견한 바 있다. 그러나 특성 0인 임의의 field에서 랭크 49보다 낮은 알고리즘을 설계하는 문제는 56년간 미해결 과제였다. AlphaEvolve는 이 문제를 해결한 최초의 방법으로, 두 개의 4×4 복소수 행렬을 곱하기 위한 랭크 48 알고리즘을 찾았다. AlphaEvolve는 초기 프로그램에 중요한 변화를 가하며, 점점 더 뛰어난 알고리즘을 설계하기 위해 여러 독창적인 아이디어들을 도입한다. 일부 파라미터 설정에 대해서는 초기 프로그램에 연구자의 아이디어를 반영했을 때(예: 평가 함수에 확률론을 추가하거나 진화적 접근법을 적용하는 경우) 성능이 더욱 향상되었다. 이러한 결과는 연구자와 AlphaEvolve 간의 과학적 협업 가능성이 중요하다는 것을 보여준다.

3.2 다양한 미해결 수학 문제에 특화된 탐색 알고리즘 발견

수학 연구의 중요한 영역 중 하나는 어떤 측정 기준에 따라 최적 또는 근접 최적의 특성을 갖는 객체나 구성을 발견하는 것이다. 예를 들어, 기하학적 도형의 조밀한 배치를 찾는 문제, 또는 특정 조합론적 또는 해석학적 제약 조건을 만족하는 함수나 집합을 식별하는 문제 등이 있다. 이러한 문제에서의 진전은 종종 기존의 모든 예제를 능가하는 단 하나의 구성을 발견하는 데 달려 있으며, 이를 통해 최적값에 대한 새로운 하한 또는 상한을 설정할 수 있다. AlphaEvolve가 이러한 문제들에 내재된 방대한 탐색 공간을 효과적으로 탐험할 수 있는 강력한 도구이며, 다양한 미해결 수학 문제를 성공적으로 해결해 냈음을 보여준다.

AlphaEvolve의 능력을 평가하기 위해, 해석학, 조합론, 수론, 기하학을 포함한 다섯 개 이상의 수학 분야에 걸친 50개 이상의 선별된 수학 문제에 AlphaEvolve를 적용했다. 이 문제들은 다양한 구체적인 매개변수 설정(예: 차원이나 크기 등)을 기준으로 평가되었다. 그 결과, 전체의 75%에서 AlphaEvolve는 현재까지 알려진 최적의 구성을 재발견했고, 20%에서는 기존의 최고 성과를 능가하는 새로운 객체를 발견, SOTA를 향상시켰다. 이 모든 경우에서 초기 출발점은 단순하거나 무작위적인 구성이었다. 이러한 결과는 AlphaEvolve가 수학 연구에 있어 매우 폭넓고 유연한 도구로서의 잠재력을 지니고 있음을 잘 보여준다.

사용된 AlphaEvolve 설정의 주요 이점 중 하나는 그 범용성과 적용 속도이다. 핵심 방법론은 휴리스틱 탐색 프로그램을 진화시키는 것에 중점을 두고 있으며, 이는 다양한 수학적 구성 문제나 수학적 추측에 대해 빠르게 적용될 수 있다. 특히 전통적인 맞춤형 접근 방식에 비해 문제에 특화된 전문가의 초기 개입이 덜 필요한 경우가 많다. 물론 깊은 수학적 통찰력은 문제 공식화나 탐색 공간 정의에 도움을 주지만, AlphaEvolve는 종종 문제 지형 내의 미묘한 구조를 파악하여 효과적인 탐색 패턴과 공략 전략을 스스로 발견하는 능력을 보여준다. 이러한 특성은 AlphaEvolve가 다양한 문제에 대해 효율적이고 대규모의 탐색을 수행하도록 해준다.

이러한 발견을 가능하게 한 핵심 방법론적 혁신은, AlphaEvolve가 대상 구조물 자체를 직접 진화시키는 것이 아니라, 휴리스틱 탐색 알고리즘을 진화시킨다는 점이다. 많은 문제들—특히 수학 분야에서 흔히 볼 수 있는, 빠르게 목표 함수를 평가할 수 있는 문제들—에 대해 반복적인 정제 전략을 사용했다. AlphaEvolve의 각 세대는 탐색 휴리스틱을 나타내는 프로그램을 진화시키는 임무를 수행한다. 이 프로그램은 일정한 시간 예산(예: 1000초)을 부여받으며, 이전 세대의 최상급 휴리스틱이 찾아낸 최선의 구조물을 입력으로 받는다. 그 목표는 이 출발점을 바탕으로, 주어진 시간 안에 더 나은 구조물을 찾는 것이다. 이러한 진화적 프로세스는 이미 고품질인 해를 더욱 개선할 수 있는 휴리스틱을 선별한다. 최종적으로 얻어진 구조물들은 종종 AlphaEvolve가 발견한 여러 단계별 특화 휴리스틱들의 연쇄적 적용의 결과였다—초기에

는 무작위나 단순한 초기 상태에서 큰 향상을 만들어내는 휴리스틱, 이후에는 거의 최적화된 구성을 정밀하게 조정하는 휴리스틱이 활용된다. 이처럼 자동으로 발견되는 다단계 적응형 탐색 전략은 수작업으로 구현하기 매우 어려우며, SOTA를 능가하는 데 결정적인 역할을 했다.

다음은 AlphaEvolve가 새로운 결과를 도출한 일부 문제들에 대한 설명이다.

해석학

- 자기상관 부등식: 여러 자기상관 부등식에서 기존의 최적 경계를 개선하는 데 성공.
- 불확정성 원리: 푸리에 해석에서 유래한 문제에 대해 기존 불확정성 원리 구성을 다듬어 약간 더 나은 상한을 제공하는 정제된 구성을 생성.

조합론 및 정수론

- Erdős의 최소 겹침 문제: 기존 최고 기록을 소폭 개선하는 새로운 상한을 수립.

기하학 및 포장 문제

- Kissing number 문제: 11차원 공간에서 kissing number의 하한을 개선. 중앙의 단위 구에 동시에 접할 수 있는 593개의 겹치지 않는 단위 구 구성을 찾아, 종전 기록인 592를 능가
- 포장 문제: 다음과 같은 다양한 포장 문제에서 새로운 결과를 달성.

• 어떤 도형 안에 N개의 점을 배치하여 최대 거리와 최소 거리의 비율을 최소화하는 문제.
• 다양한 다각형을 다른 다각형 안에 가장 효율적으로 포장하는 문제.
• 작은 면적의 삼각형 생성을 피하면서 점 집합을 배치하는 Heilbronn 문제의 변형들.

3.3 구글의 컴퓨팅 생태계 최적화

AlphaEvolve가 mission-critical infrastructure의 성능을 향상시키고 현실 세계의 성과를 이끌어내는 데 어떻게 사용되었는지를 보여준다.

3.3.1 데이터 센터 스케줄링 개선

컴퓨팅 작업을 클러스터 내 여러 기계에 효율적으로 스케줄링하는 것은 구글 데이터 센터에서 매우 중요한 최적화 문제이다. 이 작업은 자원 요구량과 기계의 용량을 고려하여 작업을 적절히 기계에 할당하는 과정을 포함한다. 비효율적인 할당은 고립된 자원을 발생시킬 수 있다. 예를 들어, 한 기계가 특정 자원(예: 메모리)은 다 소진했지만, 다른 자원(예: CPU)은 아직 남아 있어도 더 이상 작업을 받을 수 없는 상황이다. 스케줄링 효율의 향상은 이러한 고립된 자원을 회수하여 동일한 컴퓨팅 자원 내에서 더 많은 작업을 완료할 수 있게 한다. 이는 컴퓨팅 수요가 증가하더라도 자원 소비가 비례하여 늘어나지 않도록 하는 데 필수적이다. 게다가 이 문제는 디버깅 용이성과 대규모 처리 같은 전형적인 엔지니어링 난제와, 고전적으로 어려운 빈 패킹 문제가 결합되어 있어 해결이 매우 어렵다.

이 문제를 두 가지 변수로 이루어진 벡터 빈 패킹 문제로 재구성하여 해결한다. 기계들은 CPU와 메모리에 대한 용량이 정의된 빈을 나타내고, 들어오는 작업들은 특정 자원 요구량을 가진 아이템으로 간주된다. 휴리스틱 함수는 대기 중인 작업의 CPU 및 메모리 요구량과 잠재적 기계의 CPU 및 메모리 가용량을 입력으로 받아, 해당 기계에 대한 우선순위 점수를 출력한다. Borg 스케줄러는 이 우선순위 점수를 기준으로, 대기 중인 작업을 실행할 수 있고 사용 가능한 기계들 중에서 점수가 가장 높은 기계에 작업을 할당한다. 이 휴리스틱은 Borg가 이미 각 대기 작업을 실행할 수 있다고 판단한 기계들의 순위에만 영향을 미치기 때문에, 결과적으로 스케줄링 결정은 설계상 정확하게 이루어진다.

AlphaEvolve의 초기 버전은 기존에 운영 중인 휴리스틱 함수에서 진화하여, 매우 간단하면서도 효과적인 휴리스틱 함수를 발견했다. 구글 데이터 센터의 과거 작업 부하 및 용량 스냅샷을 기반으로 피드백을 제공하는 데이터 센터 시뮬레이터를 사용하여 AlphaEvolve를 학습시켰다. AlphaEvolve가 만든 휴리스틱 함수의 성능은 최근 작업 부하와 용량으로 구성된 미지의 테스트 데이터셋에서 평가되어 일반화 가능성을 확인했다. AlphaEvolve의 휴리스틱 함수가 기존 운영 중인 함수보다 우수한 성능을 보임을 확인한 후, 이를 구글 전체 클러스터에 배포했다. 배포 후 구글 전역 클러스터에서 측정한 결과 시뮬레이터 결과와 일치했으며, 이 휴리스틱 함수는 평균적으로 구글 전역 컴퓨팅 자원의 0.7%를 지속적으로 회수하는 것으로 나타났다. AlphaEvolve가 심층 강화학습 접근법 대신 선택된 이유는, AlphaEvolve의 코드 솔루션이 더 나은 성능을 제공할 뿐만 아니라 해석 가능성, 디버깅 용이성, 예측 가능성, 배포의 간편성이라는 미션 크리티컬 시스템에 필수적인 특성을 명확히 갖추고 있기 때문이다.

3.3.2 Gemini 커널 엔지니어링 향상

Gemini와 같은 대규모 모델을 학습시키는 데는 상당한 계산 자원이 필요하다. Gemini는 JAX 위에 구축되었으며, Pallas는 하드웨어 가속기에서 최적의 실행을 위해 맞추어진, 고도로 특화된 프로그램(커널)을 작성할 수 있게 하는 JAX의 확장 기능이다. 따라서 효율적인 Pallas 커널은 Gemini 학습 성능 최적화에 매우 중요하다. 커널 최적화의 핵심 요소 중 하나는 행렬 곱셈 연산에 대한 타일링 전략 조정이다. 이 기법은 큰 행렬 곱셈 연산을 더 작은 하위 문제로 나누어, 계산과 데이터 이동 간의 균형을 맞추어 전체 계산 속도를 높이는 데 핵심적 역할을 한다. 전통적으로 커널 엔지니어들은 다양한 입력 형태에 대해 거의 최적의 타일링 구성을 찾기 위해 검색 기반 자동 튜닝 또는 수동으로 작성한 휴리스틱에 의존해 왔다. 검색 기반 튜닝은 연구 흐름을 방해하며, 입력 형태가 바뀔 때마다 다시 튜닝해야 하는 번거로움이 있다. 반면, 효과적인 타일링 휴리스틱을 수동으로 작성하는 것은 그 복잡성 때문에 주요 엔지니어링 병목 현상으로 작용하며, 커널 기능과 하드웨어 세부 사항 모두에 대한 깊은 이해를 요구한다. 성능 좋은 휴리스틱의 주요 장점은 임의의 입력 형태에 대해 높은 성능을 제공할 수 있다는 점이다. 따라서 신흥 하드웨어에 대한 고성능 커널 설계 속도를 높이고, 모델 개발자가 이를 더 쉽게 활용할 수 있도록, 휴리스틱 생성 과정을 지원하는 것을 목표로 하고 있다.

이 문제를 해결하기 위해 AlphaEvolve를 활용하여 Gemini 학습에 사용되는 중요한 행렬 곱셈 커널의 타일링 휴리스틱을 최적화한다. 목표는 커널의 실제 실행 시간을 최소화하는 것이다. AlphaEvolve는 후보 코드를 제안하며 이 커널의 타일링 휴리스틱을 반복적으로 탐색하고 개선하여, 실제 TPU 가속기에서 다양한 입력 형태에 대해 실행 시간을 최소화하는 것을 목표로 한다. AlphaEvolve가 이 커널의 근본적인 수학적 연산을 변경하는 것이 아니라 타일링 전략만을 최적화하기 때문에 커널의 정확성은 구조적으로 유지된다. AlphaEvolve의 학습 및 평가 데이터셋을 구축하기 위해, 커널 사용자로부터 현실적인 커널 입력 형태를 자동으로 수집한다. 이 중 절반은 진화 과정 중 최적화 대상이 되는 학습 세트로 사용되며, 나머지 절반은 결과 휴리스틱의 일반적 적용 가능성을 시험하는 평가 세트로 사용된다.

이 자동화된 접근 방식 덕분에 AlphaEvolve는 기존 전문가가 설계한 휴리스틱 대비 모든 커널에서 평균 23%의 커널 속도 향상을 이끌어내었으며, 이에 따라 Gemini 전체 학습 시간도 1% 단축되었다. 또한 AlphaEvolve를 활용함으로써 커널 최적화에 소요되는 시간이 수개월의 전담 엔지니어링 작업에서 단 며칠의 자동화된 실험으로 크게 단축되었다. 이러한 가속화는

최적화된 커널의 배포 속도를 높여, 커널 엔지니어들이 보다 전략적이고 고차원적인 최적화 문제에 집중할 수 있게 한다. 더 나아가 AlphaEvolve는 수동 튜닝 과정을 자동화하고 Gemini 커널 사용의 작업 효율성을 향상시키는 길을 제시한다. AlphaEvolve가 발견한 타일링 휴리스틱은 실제 서비스에 배포되어 Gemini 학습 효율성과 Gemini 팀의 연구 및 엔지니어링 속도를 직접 향상시키고 있다. 이는 AlphaEvolve의 역량을 통해 Gemini가 자신의 학습 프로세스를 스스로 최적화하는 새로운 사례이다.

3.3.3 하드웨어 회로 설계 지원

Google의 TPU와 같은 특수 하드웨어는 대규모 AI 시스템을 효율적으로 운영하는 데 필수적이다. 그러나 새로운 컴퓨터 칩을 설계하는 과정은 매우 복잡하고 시간이 많이 소요되며, 종종 수년에 걸쳐 진행된다. 이 과정에서 중요한 단계인 Register Transfer Level(RTL) 최적화는 전력, 성능, 면적 등 여러 지표를 개선하기 위해 하드웨어 설계 코드를 수동으로 다시 작성하는 작업을 포함하며, 숙련된 엔지니어들의 수개월에 걸친 반복적 노력이 필요하다.

이번 작업에서 AlphaEvolve는 행렬 곱셈 유닛 내 핵심 TPU 산술 회로의 최적화된 Verilog 구현을 최적화하는 과제를 맡았다. 최적화 목표는 구성 요소의 핵심 기능을 유지하면서 면적과 전력 소비를 모두 줄이는 것이었다. 무엇보다도, 최종 제안은 수정된 회로가 기능적 정확성을 유지함을 확인하는 강력한 검증 방법을 통과해야 한다. AlphaEvolve는 불필요한 비트를 제거하는 간단한 코드 재작성 방안을 찾아냈으며, 이 변경 사항은 TPU 설계자들에 의해 정확성이 검증되었다. 비록 이 특정 개선 사항이 후속 합성 도구에 의해서도 독립적으로 발견되었지만, AlphaEvolve가 RTL 단계에서 소스 RTL을 다듬고 설계 흐름 초기에 최적화를 제공할 수 있음을 보여준 중요한 사례이다.

차기 TPU에 통합된 이 개선은 AlphaEvolve를 통해 이루어진 Gemini의 TPU 산술 회로에 대한 직접적인 첫 번째 기여 사례로서, 앞으로의 추가 기여를 위한 길을 열었다. AlphaEvolve의 주요 장점 중 하나는 하드웨어 엔지니어들이 사용하는 표준 언어인 Verilog로 제안된 변경 사항을 직접 전달하여 신뢰성을 높이고 도입을 용이하게 한다는 점이다. 이 초기 탐구는 LLM 기반 코드 진화가 하드웨어 설계를 지원하여 시장 출시 시간을 단축할 수 있다는 것을 보여준다.

3.3.4 컴파일러가 생성한 코드를 직접 최적화하기

트랜스포머 아키텍처는 LLM부터 AlphaFold에 이르기까지 대부분의 현대 신경망에서 사용된다. 트랜스포머의 핵심 연산은 어텐션 메커니즘이며, 이는 주로 FlashAttention을 사용해 구현된다. 구글 딥마인드의 스택에서는 FlashAttention이 Pallas에서 가속기 커널로 구현되어 있으며, 입력 준비와 출력 후처리를 담당하는 상위 수준의 JAX 코드로 감싸져 있다. 그 후 머신러닝 컴파일러인 XLA가 이 구현을 특정 하드웨어에서 실행할 수 있도록 점점 더 상세한 중간 표현(IR) 시퀀스로 변환한다. 이 단계들에서 메모리 접근 조정이나 연산 스케줄링에 대한 개선된 결정은 특정 하드웨어에서 실행 시간을 크게 단축할 수 있다.

AlphaEvolve에게 FlashAttention 커널과 전후처리 코드를 포함하는 XLA가 생성한 IR(중간 표현)을 직접 최적화하는 과제를 부여했다. GPU에서 대규모 추론에 사용되는 매우 영향력 있는 트랜스포머 모델에 해당하는 구성을 최적화하여, 모듈의 전체 실행 시간을 최소화하는 것을 목표로 했다. 이 작업은 특히 어려웠는데, 그 이유는 (1) IR이 개발자가 직접 수정하기보다

는 디버깅 용도로 설계되었고, (2) 컴파일러가 생성했으며 이미 매우 최적화되어 있기 때문이다. AlphaEvolve가 제안한 각 수정 사항은 최적화 과정 내내 수치적 정확성을 보장하기 위해 무작위 입력에 대해 기준(수정 전) 코드와 대조하여 검증되었다. 최종 코드 버전은 모든 가능한 입력에 대해 인간 전문가들에 의해 엄격하게 올바름이 확인되었다.

AlphaEvolve는 IR이 노출하는 두 가지 추상화 수준 모두에서 의미 있는 최적화를 제공할 수 있었다. 첫째, 관심 있는 구성의 FlashAttention 커널은 32% 속도 향상을 이루었다. 둘째, AlphaEvolve는 커널 입력과 출력의 전처리/후처리 과정에서 개선점을 찾아 이 부분에서 15%의 속도 향상을 달성했다. 이 결과들은 AlphaEvolve가 컴파일러가 생성한 코드를 최적화할 수 있음을 보여주며, 발견된 최적화를 특정 사용 사례를 위한 기존 컴파일러에 통합하거나, 장기적으로는 AlphaEvolve 자체를 컴파일러 워크플로우에 포함시킬 가능성을 제시한다.

4 소거 실험 (Ablations)

더 빠른 행렬 곱셈을 위한 텐서 분해 찾기(3.1)와 kissing number의 하한 계산(3.2)에 대해 소거 실험을 하였다. 이는 AlphaEvolve의 구성 요소들의 효과를 이해하기 위한 목적이다.

·**진화적 접근법**: AlphaEvolve는 이전에 생성된 프로그램들을 데이터베이스에 저장하고, 이후 반복에서 더 나은 프로그램을 얻기 위해 이를 활용하는 진화적 접근법을 사용한다. 진화의 중요성을 분석하기 위해, 동일한 초기 프로그램을 언어 모델에 반복적으로 입력하는 대안적 접근 방식 “진화 없음(No evolution)”을 고려했다.

·**프롬프트 내 문맥:** AlphaEvolve는 대형 콘텍스트 윈도우를 가진 강력한 언어 모델을 사용하며, 문제에 특화된 문맥 정보를 프롬프트에 제공함으로써 출력 품질을 크게 향상시킬 수 있다. 문맥의 중요성을 테스트하기 위해, 프롬프트에 명시적인 문맥을 추가하지 않는 대안적 접근 방식 “프롬프트 내 문맥 없음(No context in the prompt)”을 고려했다.

·**메타프롬프트:** AlphaEvolve는 언어 모델에 제공되는 프롬프트를 개선하기 위해 메타프롬프트를 사용한다. 이를 통해 인간이 작성한 프롬프트로 얻을 수 있는 성능을 능가할 수도 있다. 메타프롬프트의 효과를 테스트하기 위해, 텐서 분해 작업에서 이를 비활성화하는 접근 방식 “메타프롬프트 진화 없음(No metaprompt evolution)”을 고려하였다.

·**전체 파일 진화:** 이전의 FunSearch 같은 접근 방식과는 달리, AlphaEvolve는 단일 함수에 집중하는 대신 전체 코드베이스를 진화시킬 수 있다. 전체 파일 진화의 중요성을 테스트하기 위해, 텐서 분해 문맥에서 손실 함수만을 진화시키는 대안적 방식 “전체 파일 진화 없음(No full-file evolution)”을 고려하였다.

·**강력한 언어 모델:** AlphaEvolve는 다양한 샘플을 얻기 위해 소형 및 대형 언어 모델의 조합에 의존한다. 이 구성 요소의 중요성을 이해하기 위해, 단일 소형 기본 모델만 사용하는 대안적 접근 방식 “소형 기본 LLM만 사용(Small base LLM only)” 을 고려하였다.

5. 관련 연구 (Related Work)

진화적 방법: AlphaEvolve는 반복적으로 돌연변이 및 교차 연산자를 사용하여 프로그램 집합을 진화시키는 진화적 혹은 유전자 알고리즘에 대한 오랜 연구를 확장한 것이다. 특히, 고전적인 진화 기법은 기호 회귀(symbolic regression), 과학 및 알고리즘 발견 자동화, 스케줄링 문제 등에 성공적으로 활용되어 왔다. 이러한 방법들의 과제는 직접 설계한 진화 연산자를 사용하는 데 있는데, 이는 설계가 어렵고 도메인의 중요한 속성을 반영하지 못할 수 있다. 이에 반해, AlphaEvolve는 LLM을 활용해 이러한 연산자의 생성을 자동화한다. 즉, 미리 정의된

돌연변이 연산자 없이도 LLM이 가진 세계 지식을 바탕으로 프로그램을 변형할 수 있다.

AlphaEvolve는 최근 등장한 LLM과 진화 기법의 결합 시도들을 활용하고 있으며, 특히 수학적 발견을 위한 접근 방식으로 Romera-Paredes 등이 제안한 FunSearch 시스템을 확장한 것이다. FunSearch는 이후 베이지안 최적화를 위한 획득 함수 학습, 인지 모델 발견, 그래프 간 거리 계산, 조합적 경쟁 프로그래밍 등의 작업에 사용되었다.

AlphaEvolve는 FunSearch 및 그 최신 구현을 다음 세 가지 측면에서 뛰어넘는다: ① FunSearch는 단일 Python 함수만 진화시킬 수 있었지만, AlphaEvolve는 다양한 언어로 작성된 전체 코드베이스의 진화가 가능하다. ② FunSearch는 단일 목적 함수만 최적화할 수 있었지만, AlphaEvolve는 다목적 최적화(multiobjective optimization)를 지원한다. ③ FunSearch는 비교적 작은 모델만 사용하고 코드 학습에만 특화되어 있었지만, AlphaEvolve는 최신 대형 LLM과 풍부한 자연어 기반의 문맥 및 피드백을 활용한다. 이러한 확장을 통해 AlphaEvolve는 FunSearch가 다루기 어려웠던 복잡한 문제들을 해결할 수 있게 되었다.

또한 유사한 접근으로는 Lehman et al.의 시뮬레이션 로봇을 위한 프로그래밍 정책 발견, Hemberg et al.의 코드 합성 방법이 있으며, 이 밖에도 기호 회귀, 조합 최적화 휴리스틱 발견, 분자 구조 합성 등의 다양한 과학적, 수학적 작업에 유사한 방법이 사용되어 왔다. 또한, LLM 프롬프트 개선, 신경망 구조 탐색 등에도 LLM 기반 진화가 활용되었다. AlphaEvolve는 적용 범위, 유연성, 확장성 면에서 이러한 기존 접근들과 차별화된다.

몇몇 최근 연구는 LLM 기반 진화 개념에 보완 아이디어를 추가하기도 했다. 예를 들어, Surina et al.는 강화 학습 기반 미세 조정을 결합하고, Grayeli et al.는 우수한 프로그램을 요약해 자연어로 개념화하는 학습 단계를 도입했다. 이러한 보완 기법들이 AlphaEvolve 수준의 대규모 시스템에서 어떤 이점을 제공할지는 추가 연구가 필요하다.

진화 기법은 최근 AI Co-Scientist 프로젝트에도 활용되었다. 이 시스템은 과학적 발견 자동화를 목표로, 가설 생성, 가설 평가, 문헌 검토 등의 작업을 수행하는 별도의 에이전트를 사용한다. AI Co-Scientist는 가설과 평가 기준을 자연어로 표현하는 반면, AlphaEvolve는 코드를 진화시키고 프로그램 기반의 평가 함수를 사용한다. 이로 인해 LLM의 환각 문제를 회피하고, 진화를 매우 긴 시간 동안 안정적으로 수행할 수 있다. 하지만 이 두 접근을 결합해 자연어 기반 표현과 프로그램 기반 표현을 유연하게 혼합하는 방식도 가능하다.

슈퍼 최적화 및 알고리즘 발견: AlphaEvolve는 실행 피드백을 기반으로 초기 프로그램을 반복적으로 개선한다는 점에서, 코드 슈퍼 최적화 방법으로 볼 수 있다. 코드 슈퍼 최적화는 1980년대부터 연구되어 왔으며, LLM 이전에는 체계적 열거, 유전 탐색, 몬테카를로 샘플링, 딥 강화 학습 등이 사용되었다. 특정 문제(예: 행렬 곱셈)에 국한된 경우, AlphaTensor와 같이 증명 가능한 알고리즘을 발견하는 시스템도 등장한 바 있다.

최근에는 LLM 기반의 슈퍼 최적화 및 알고리즘 발견 방법들이 활발히 연구되고 있으며, 이는 AlphaCode와 같은 프로그래밍 대회 성공 사례를 통해 그 가능성이 입증되었다. 예를 들어, LLM 에이전트를 활용하여 GPU 커널 내 attention 연산 또는 사용자 지정 연산을 최적화하기도 하며, 진화 알고리즘 자체를 새롭게 발견하거나, 언어 모델을 훈련하거나, 대규모 컴퓨팅 시스템을 최적화하는 작업에도 사용되고 있다. 또한, 여러 LLM 에이전트가 협력적으로 문제 해결에 나서는 방식도 제안되었다. 이전의 LLM 기반 알고리즘 발견 연구가 가능성을 보여준 반면, AlphaEvolve는 진화 알고리즘에 LLM을 통합하여 더 어려운 문제들을 해결할 수 있게 된 것이 큰 차별점이다.

과학 및 수학 발견을 위한 AI: 지난 10년간 AI는 단백질 구조 예측, 양자물리학, 기후 과학

등 다양한 과학 분야에 걸쳐 활용되어 왔다. 특히, 재료과학, 화학, 생물정보학, 지구과학, 양자물리 등에서는 LLM 기반 과학 문제 해결 기법들이 활발히 개발되고 있다. 이들 중 많은 방법은 LLM을 활용해 가설 생성, 랭킹, 실험 설계, 실행, 데이터 분석 등 과학 발견의 여러 단계를 자동화한다. 특히 AlphaEvolve와 관련 있는 방법으로는 LLM 기반 트리 탐색 알고리즘이나 LLM 기반 진화 알고리즘이 있다. 이 외에도 실험 설계 최적화, 실험 실행 및 워크플로우 자동화, 데이터 분석 자동화 등 다양한 응용이 있다. AlphaEvolve는 대부분의 기존 방법과 달리, 자연어가 아닌 코드로 표현된 가설 및 평가 지표를 사용한다는 점에서 차별화된다.

AI는 순수 수학 분야의 발전에도 기여하고 있으며, 그중 FunSearch는 수학 명제에 대한 반례나 증거를 찾는 도구로서 LLM 기반 진화를 강력한 접근으로 자리매김했다. 이는 수학 명제에 대한 형식적 또는 비형식적 증명을 찾는 문제와 상보적인 문제다.

6. 논의 (Discussion)

AlphaEvolve는 최첨단 LLM과 자동 평가 지표를 진화적 프레임워크 내에서 결합함으로써, 수십 년 된 수학 문제에 대한 새로운 발견뿐만 아니라 고도로 최적화된 컴퓨팅 스택에 대한 실질적인 개선을 이끌어낼 수 있는 놀라운 가능성을 보여준다.

흥미롭게도, AlphaEvolve는 동일한 문제에 대해 서로 다른 방식으로 접근하는 것을 가능하게 한다. 예를 들어, 해결책을 직접 탐색하거나, 해결책을 처음부터 생성하는 함수를 찾거나, 해결책을 찾는 탐색 알고리즘 자체를 진화시키는 방식이 있다. AlphaEvolve를 어떤 방식으로 적용하느냐에 따라 편향(bias)이 달라지며(예: 구성적 함수를 찾는 방식은 높은 대칭성을 갖는 객체를 찾는 경향이 있음), 이는 문제 유형에 따라 적합한 접근을 다르게 만든다.

AlphaEvolve는 테스트 시점에 실행되는 연산 에이전트로 볼 수 있으며, 진화 과정을 통해 기본 LLM의 능력을 반복 샘플링보다 훨씬 뛰어난 수준으로 향상시킨다. 이는 한편으로는, 기계 피드백이 테스트 시점 연산을 확장함으로써 새로운 과학적 발견이나 실질적으로 가치 있는 최적화를 가능하게 한다는 점을 강력하게 입증한 사례다.

다른 한편으로는, AlphaEvolve를 통해 향상된 기본 LLM의 성능을 차세대 LLM에 증류하는 것이 자연스러운 다음 단계가 될 것이다. 이는 그 자체로도 의미가 있으며, 차세대 AlphaEvolve의 성능을 끌어올리는 데에도 기여할 수 있다. 이러한 증류 외에도, AlphaEvolve는 자신의 기반 인프라나 미래 버전의 LLM 효율성을 개선하는 실질적인 발견을 이끌어낼 수 있다는 점에서도 흥미롭다. 현재로서는 이러한 성과가 적당한 수준이며, AlphaEvolve의 다음 버전을 개선하는 피드백 루프는 수개월 단위이다. 그러나 이러한 개선이 축적됨에 따라, 강력한 평가 함수를 갖춘 다양한 환경을 구축하는 것의 가치가 점점 더 인식될 것이고, 앞으로 더 높은 실용적 가치를 가진 발견으로 이어질 수 있다.

AlphaEvolve의 주된 한계는 자동 평가기를 설계할 수 있는 문제에 한해서만 작동한다는 점이다. 이는 수학 및 계산 과학 분야의 많은 문제에 적용 가능하지만, 자연과학과 같이 실험의 일부만 자동화 또는 시뮬레이션이 가능한 분야에서는 적용이 제한적일 수 있다. AlphaEvolve는 LLM이 제공하는 아이디어 평가를 허용하지만, 이는 본 시스템이 최적화한 설정은 아니다. 그러나 동시 진행 중인 연구들은 이러한 방식이 가능함을 보여주고 있으며, 자연스러운 다음 단계는 두 설정을 연결하는 것이다. 즉, LLM이 고차원적인 아이디어에 피드백을 제공하고, 그 후 코드 실행을 통한 기계적 피드백이 가능한 구현 단계로 넘어가는 방식이 될 수 있다.

“AlphaGenome: Advancing Regulatory Variant Effect Prediction with a Unified DNA Sequence Model”

요약: DNA 서열로부터 기능적 유전체 측정값을 예측하는 딥러닝 모델은 유전적 조절 코드를 해독하는 데 강력한 도구이다. 기존 방법들은 입력 서열의 길이와 예측 해상도 사이에서의 trade-off 때문에, 이로 인해 처리할 수 있는 모달리티의 범위와 성능이 제한된다. AlphaGenome은 1 메가 베이스 길이의 DNA 서열을 입력으로 받아, 유전자 발현, 전사 시작, 크로마틴 접근성, 히스톤 변형, 전사인자 결합, 크로마틴 접촉 지도, 스플라이스 사이트 사용, 스플라이스 접합부의 위치 및 강도 등 다양한 모달리티에 걸쳐 단일 염기쌍 수준의 해상도로 수천 개의 기능적 유전체 트랙을 예측한다. AlphaGenome은 인간과 생쥐 게놈을 기반으로 학습되었으며, 변이 효과 예측에 대한 26가지 평가 중 24가지에서 기존의 최고 외부 모델과 동등하거나 이를 능가하는 성능을 보였다. 모든 모달리티에 걸쳐 변이 효과를 동시에 정확하게 평가하는 AlphaGenome은 TAL1 온코진 근처의 임상적으로 중요한 변이 메커니즘을 충실히 재현한다. 보다 폭넓은 활용을 위해, DNA 서열로부터 유전체 트랙 및 변이 효과 예측을 수행할 수 있는 도구도 제공한다.

1 서론

게놈 서열 변이가 미치는 영향을 해석하는 일은 여전히 생물학의 주요 과제이다. 특히 단백질을 코딩하지 않는 영역에 위치한 비부호화 변이는 유발할 수 있는 분자 수준의 결과가 매우 다양하기 때문에 해석이 더욱 어렵다. 예를 들어, 비부호화 변이는 크로마틴 접근성, 후성 유전적 변형, 3차원 크로마틴 구조와 같은 게놈의 특성에 영향을 줄 수 있다. 또한, 이러한 변이는 스플라이싱(splicing) 변화를 통해 mRNA의 발현 수준을 조절하거나 서열 구성을 변경함으로써 mRNA의 가용성에도 영향을 줄 수 있다. 변이는 세포 유형이나 조직 특이적인 영향을 보이기도 한다. 인간에게서 관찰되는 유전적 변이의 98% 이상이 비부호화 영역에 위치한다는 점을 고려할 때, 이처럼 방대한 수의 변이들이 초래할 수 있는 복잡한 효과를 전반적으로 규명하는 일은 계산적 예측 없이는 사실상 불가능하다.

계산적 방법은 실험 데이터를 통해 패턴을 학습함으로써 변이의 효과를 예측하고 설명할 수 있다. 이 중 sequence-to-function(서열 기반 기능 예측) 모델이라고 불리는 한 종류의 방법은 DNA 서열을 입력으로 받아 게놈 트랙을 예측한다. 게놈 트랙은 실험을 통해 얻어진 읽힘 수(read coverage), 카운트, 또는 신호 값을 각 DNA 염기쌍에 연관시킨 데이터 형식이다. 이들은 다양한 데이터 모달리티를 포함하며, 예를 들어 다음과 같은 실험에서 측정된 유전자 발현 (RNA-seq, CAGE-seq, PRO-cap), 스플라이싱 (스플라이스 사이트, 사용량, 접합부), DNA 접근성 (DNase-seq, ATAC-seq), 히스톤 변형 (ChIP-seq), 전사인자 결합 (TF ChIP-seq), 크로마틴 구조 (Hi-C/micro-C) 등이 포함된다. 이러한 sequence-to-function 모델이 성공적으로 학습되면, 입력된 DNA 서열로부터 실제 실험 측정값을 정확히 예측할 수 있다. 나아가, 참조 서열과 변이 서열(alternative sequence)에 대해 예측된 게놈 트랙을 비교함으로써, 해당 변이가 야기하는 분자 수준의 영향을 예측할 수 있다.

현재 딥러닝 기반의 sequence-to-function 모델들은, 다양한 생물학적 조절 방식에 대한 변이의 영향을 예측하는 능력에 있어 두 가지 근본적인 트레이드 오프에 직면해 있다. 첫 번째

는 계산 자원의 한계로 인해, 모델이 장거리 게놈 상호작용을 포착하는 것과 염기 수준의 예측 해상도를 확보하는 것 사이에서 트레이드 오프를 선택해야 한다는 점이다. 예를 들어, SpliceAI, BPNet, ProCapNet, APARENT 등의 모델은 염기 단위의 고해상도 예측을 제공하지만, 입력 서열 길이가 짧게 제한되어 있어 (예: 10kb 이하), 먼 거리의 조절 요소가 미치는 영향을 놓치게 된다. 반면, Enformer나 Borzoi와 같은 모델은 더 긴 서열(약 200kb에서 500kb)을 처리하여 넓은 문맥 정보를 포착할 수 있지만, 그 대가로 출력 해상도가 낮아진다 (예: 128bp 또는 32bp 단위로 출력). 이로 인해, 스플라이스 사이트, 전사인자 결합 지점 (footprints), 폴리애데닐화 위치와 같은 미세한 조절 특징들이 흐릿해질 수 있다.

두 번째 트레이드 오프는 다양한 모달리티를 포착하는 것과 특정 모달리티에 특화되는 것 사이에 있다. 여러 SOTA 모델들은 하나의 모달리티에 매우 특화되어 있다. 예를 들어, SpliceAI는 스플라이스 사이트 예측에, ChromBPNet은 국소적인 크로마틴 접근성 예측에, Orca는 3D 게놈 구조 예측에 특화되어 있다. 그러나 이러한 특화 모델들만으로는 변이가 여러 모달리티에 걸쳐 일으키는 다양한 분자적 결과를 포괄하기에는 충분하지 않다. 예를 들어, splicing이라는 단일 모달리티 내에서도, SpliceAI나 Pangolin과 같은 특화 모델은 스플라이스 사이트 예측과 같은 일부 측면만을 다루고, 스플라이스 접합부 예측이나 스플라이스 사이트 간 경쟁과 같은 다른 중요한 측면은 다루지 않는다. 반면에, DeepSEA, Basenji, Enformer, Sei, Borzoi 등의 모델은 멀티모달 모델의 유용성과 실용성을 보여주었다. 이러한 모델들은 사용자가 여러 특화 모델을 따로 사용하는 대신, 하나의 모델로 여러 모달리티를 처리할 수 있게 해준다. 게다가 이들 모델이 학습한 general sequence representation은 새로운 작업에 대해 쉽게 파인튜닝 될 수 있도록 한다. 하지만 이러한 범용 모델들은 일부 작업 (예: 스플라이싱)에서는 특화 모델보다 성능이 떨어질 수 있으며, 경우에 따라 특정 모달리티 (예: 접촉 지도 contact maps) 자체가 빠져 있을 수도 있다.

AlphaGenome은 멀티모달 예측, 긴 서열 문맥, 염기쌍 수준의 해상도를 하나의 통합된 프레임워크에 결합한 모델이다. AlphaGenome은 1 메가 베이스(Mb) 길이의 DNA 서열을 입력으로 받아, 다양한 세포 유형에 걸쳐 폭넓은 범위의 게놈 트랙을 예측한다. AlphaGenome의 스플라이싱 예측에는 스플라이스 사이트 사용량 예측과 더불어, 새로운 방식의 스플라이스 접합부 예측도 포함된다. AlphaGenome의 성능을 종합적인 벤치마크를 통해 평가했으며, 이전에 보지 못한 DNA 서열에 대한 게놈 트랙 예측 정확도뿐만 아니라, 변이 효과 예측 과제에 대한 모델의 효과성도 포함된다. 그 결과, AlphaGenome은 24개의 게놈 트랙 예측 과제 중 22개, 변이 효과 예측 과제 중 26개 중 24개에서 SOTA 성능을 달성했다. 또한, AlphaGenome의 성능을 설명하고 향후 sequence-to-function 모델 설계를 위한 인사이트를 제공하기 위해, 출력 해상도, 서열 길이, distillation, 모달리티 조합에 대한 광범위한 어블레이션 실험 (ablation studies)을 수행했다. AlphaGenome이 게놈 내 조절 코드를 분석하기 위한 강력하고 확장 가능한 기반이 될 것으로 기대한다.

2 결과

2.1 DNA 서열 기반 기능 예측 통합 모델

AlphaGenome은 인간과 생쥐의 DNA로부터 다양한 분자 표현형의 서열 기반을 학습하도록 설계된 딥러닝 모델이다. 이 모델은 유전자 발현(RNA-seq, CAGE, PRO-cap), 정교한 스플라이싱 패턴(스플라이스 사이트, 사용량, 접합부), 크로마틴 상태(DNase, ATAC-seq, 히스톤 변

형, 전사인자 결합), 크로마틴 접촉 지도 등 11개 모달리티에 걸쳐 5,930개의 인간 게놈 트랙 또는 1,128개의 생쥐 게놈 트랙을 동시에 예측한다. 이러한 예측은 다양한 생물학적 맥락—예: 여러 조직 유형, 세포 유형, 세포주—을 포함하고 있다. AlphaGenome은 1 메가 베이스(Mb) 길이의 DNA 서열을 기반으로 이러한 예측을 수행한다. 이 콘텍스트 길이는 관련된 장거리 조절 영역의 상당 부분을 포함하도록 설계되었다. 실제로 검증된 인핸서-유전자(enhancer-gene) 쌍 중 99% (465/471)가 1Mb 이내에 존재하는 것으로 보고되었다.

AlphaGenome은 U-Net에서 착안한 아키텍처를 활용하여, 입력 DNA 서열을 효율적으로 처리해 두 가지 형태의 서열 표현으로 변환한다. 하나는 1차원 임베딩으로, 염기쌍 단위(1bp)와 128bp 해상도로 생성되며, 선형 게놈 구조를 나타낸다. 다른 하나는 2차원 임베딩으로, 2048bp 해상도로 구성되며, 게놈 구간 간의 공간적 상호작용을 표현한다. 1차원 임베딩은 게놈 트랙 예측의 기반으로 사용되고, 2차원 임베딩은 쌍별 상호작용(contact maps) 예측의 기반이 된다. 이 아키텍처 내에서는 합성곱 층이 정밀한 예측을 위해 필요한 국소적인 서열 패턴을 모델링하며, 트랜스포머 블록은 인핸서-프로모터 간 상호작용과 같이 보다 거칠지만 장거리 의존성을 모델링한다. 1Mb 전체 서열에 대해 염기쌍 수준으로 학습하는 것은 8개의 상호 연결된 TPUv3 디바이스에 걸친 서열 병렬화를 통해 구현된다. 게놈 트랙 예측은 이러한 서열 임베딩에 선형 변환을 적용함으로써 수행되며, 스플라이스 접합부 카운트 예측의 경우에는 도너/엑셉터 쌍 간의 1차원 임베딩 상호작용을 포착하는 별도의 메커니즘이 사용된다.

AlphaGenome을 사전 학습과 distillation의 두 단계로 나누어 학습시켰다. 사전 학습 단계에서는 실험 데이터를 활용해 두 가지 유형의 모델을 생성한다. 먼저, 폴드별(fold-specific) 모델은 4-폴드 교차 검증 방식을 사용하여 학습된다. 즉, 기준(reference) 게놈의 4분의 3을 학습에 사용하고, 나머지 4분의 1은 검증 및 테스트용으로 한다. 이 모델들은 AlphaGenome의 일반화 성능을 평가하기 위해, 보지 못한 테스트 구간의 게놈 트랙 예측에 사용된다. 또 다른 유형은 전체 폴드 모델로, 기준 게놈의 모든 구간을 학습에 사용한다. 이 모델들은 두 번째 단계인 디스틸레이션의 교사 모델 역할을 한다. 디스틸레이션 단계에서는, 사전 학습된 아키텍처와 동일한 구조를 공유하는 하나의 학생 모델이, 무작위로 증강된 입력 서열을 바탕으로, 모든 폴드 교사 모델의 앙상블 출력을 예측하도록 학습된다. 이러한 디스틸된 학생 모델은 강인성과 변이 효과 예측(VEP) 정확도 측면에서 개선된 성능을 단일 모델 인스턴스로 달성한다. 이 모델은 하나의 변이에 대해 한 번의 호출만으로, 모델링된 모든 모달리티와 세포 유형에 걸친 예측을 수행할 수 있다. 또한, NVIDIA H100 GPU에서 1초 미만의 시간으로 예측을 수행할 수 있어, 여러 개의 독립적으로 학습된 모델을 앙상블하는 기존 방식에 비해 대규모 변이 예측 작업에서 매우 효율적이다.

2.2 모델 상세 설명

AlphaGenome은 1Mb의 DNA 서열과 종 식별자를 입력으로 받아 다양한 유전체 특징을 예측하는 딥러닝 기반 모델이다. 이 모델의 출력은 여러 해상도에서 크로마틴 접근성, 전사인자 결합, 유전자 발현 등과 같은 신호를 나타내는 1D 트랙, 2D 염색체 접촉 지도, 그리고 스플라이싱 관련 예측(예: 스플라이스 사이트 확률, 스플라이스 사이트 사용률(SSU), 스플라이스 접합 수치) 등을 포함한다. AlphaGenome은 Enformer와 Borzoi 등의 모델에서 착안하여 컨볼루션 및 트랜스포머 기반의 인코더-디코더 구조를 채택하고 있으며, 인코더의 잔차 연결을 포함하는 U-Net 스타일의 디코더 등 여러 핵심적인 아키텍처 특징을 도입했다. 이 모델은 다

음의 다섯 가지 핵심 구성 요소를 통해 정보를 처리한다: 시퀀스(서열) 인코더, 트랜스포머 타워, 쌍별 상호작용 블록, 서열 디코더, 작업(task)별 출력 헤드. AlphaGenome은 약 4억 5천만 개의 학습 가능한 파라미터를 다음과 같이 가진다: 인코더 20%, 서열 트랜스포머 28%, 쌍별 상호작용 블록 15%, 디코더 25%, 출력 임베딩 및 예측 헤드 12%.

2.2.1 시퀀스 인코더(Sequence Encoder)

시퀀스 인코더는 입력된 DNA 서열을 1bp 해상도(길이 1Mb 또는 2^{20} bp)에서 시작하여, 128bp 해상도의 임베딩(길이 8192 또는 2^{13})으로 점진적으로 다운 샘플링한다. 이 과정은 총 7단계에 걸쳐 이루어진다. 특징 채널의 수는 표현력을 높이기 위해 768개에서 1536개까지 증가한다. 이러한 다운 샘플링은 컨볼루션 블록을 반복 적용한 후, 각 블록마다 스트라이드가 2인 맥스 풀링을 수행으로 이루어진다.

인코더의 핵심 구성 요소는 conv_block이며, 이 블록은 다음 연산들을 순차적으로 수행한다: RMSBatchNorm(일반적인 배치 정규화와 유사하지만, 샘플 평균으로 이동하지 않고, 학습된 스케일과 옵셋만을 적용), GeLU75 활성화 함수, 표준화된 1D 컨볼루션(커널 너비는 5이며, 스케일된 가중치 정규화를 사용). 이 방법은 가중치를 표준화하고 스케일링하여 재파라미터화함으로써, 활성화 값의 안정적인 크기 유지에 효과적임을 확인했다. 또한, 각 RMSBatchNorm 층은 학습 중에 채널별 분산에 대한 지수 이동 평균(EMA, 감쇠율 0.9)을 유지하며, 이 EMA는 추론 시 정규화에 사용된다. 각 컨볼루션 블록에는 잔차 연결이 추가된다. 잔차 연결 시 채널 차원이 서로 다를 경우(예: downres_block에서처럼), 입력에 0으로 패딩을 추가한 후 덧셈 연산을 수행한다. 모든 컨볼루션 층은 "same" 패딩을 사용한다.

2.2.2 트랜스포머 타워

인코더 이후에는, Transformer 타워가 128bp 해상도의 시퀀스 임베딩을 처리하여, 전체 1Mb 입력 컨텍스트 전반에 걸친 장거리 상호작용을 모델링한다. 이 타워는 9개의 Transformer 블록을 쌓은 구조로 이루어져 있다. 각 블록은 두 구성 요소(Multi-Head Attention 층, two layer MLP 층)로 이루어져 있으며, 모두 잔차 연결되어 있다.

모든 두 번째 MHA 블록 직전에, 2048bp 해상도에서의 쌍별 표현이 시퀀스 임베딩을 기반으로 초기화 또는 갱신된다, 이 쌍별 표현은 주로 접촉 지도 예측에 사용되며, 동시에 MHA 레이어에 추가되는 bias 항으로도 활용된다. attention_bias_block 함수는 쌍별 채널(F)을 어텐션 헤드 수(8개)에 맞게 투영한 뒤, 두 시퀀스 축에 걸쳐 16번(= 2048bp ÷ 128bp) 반복하여 전체 시퀀스 길이 차원(8192)에 맞춘다. 이 bias는 softmax 이전에 attention logits에 직접 더해지며, 이는 AlphaFold2에서 사용된 기법과 유사하다.

MLP 블록은 표준 트랜스포머 설계를 따르며, ReLU 활성화 함수를 사용하고, 은닉층에서 채널 수를 2배로 확장한다. 드롭아웃은 MHA 및 MLP 블록 내에서 훈련 중에만 적용되며, 드롭아웃 비율은 사전학습 시 0.3, 지식 증류(distillation) 시 0.1이다.

2.2.3 쌍별 상호작용 블록

트랜스포머 타워 블록과 교차적으로 배치되어 있으며, pair_update_block 모듈은 트랜스포머 블록 0, 2, 4, 6, 8번에서 MHA 층 이전에 실행되어 서열 위치 간의 쌍별 상호작용을 나타내는 2차원 텐서를 조작하고 업데이트한다. 이 쌍별 표현은 2048bp 해상도를 가지며 128개의

채널로 구성된다. 이는 주로 접촉 지도 예측에 사용되며, 동시에 트랜스포머의 어텐션 메커니즘에 바이어스 항으로 기여한다. 각 pair_update_block은 다음 단계를 따른다: 먼저 sequence_to_pair_block을 통해 서열 임베딩에서 업데이트를 계산하고, 이 업데이트는 기존의 pairwise 상태를 초기화하거나 기존 값에 더해진다. 그 후, 두 단계의 잔차 기반 정제 과정 (row_attention_block 및 pair_mlp_block)을 수행한다.

sequence_to_pair_block 함수는 먼저 128bp 해상도의 서열 임베딩(S = 8192)을 평균 풀링을 통해 2048bp 해상도로 다운 샘플링한다. 그 후, 상대 위치 인코딩이 강화된 일종의 어텐션 메커니즘을 사용하여 쌍별 특성을 계산한다. 위치 인코딩은 central_mask_features 함수에 의해 생성되며, 32개의 지수적으로 간격이 벌어진 거리 임계값을 기반으로 한다. 그 다음, 상대 거리의 방향성을 표현하기 위해 이 32개의 특성에 상대 거리의 부호를 곱한 복사본을 이어 붙여 총 64채널의 위치 표현을 만든다. 이 상대 어텐션 항은 Transformer-XL에서 제안된 relative_shift 연산을 통해 효율적으로 계산된다. 최종적인 sequence-to-pair 표현은 다음 과정을 통해 얻어진다: 쿼리-키와 상대 위치 항을 결합하여 투영 서열 임베딩으로부터 파생된 투영의 외적 합을 더함, 드롭아웃 적용.

쌍별 임베딩은 두 개의 트랜스포머 스타일 연산을 통해 잔차 연결 내에서 추가로 정제된다. row_attention_block은 자기 어텐션 메커니즘을 구현하는데, 이는 PxP 쌍별 행렬의 두 번째 축을 따라 작동한다. 즉, 각 위치 (i, j)는 같은 행 i에 있는 모든 위치 (i, k)에 주의를 기울이게 된다. 열 방향으로 작동하는 두 번째 어텐션 블록을 적용했을 때, 후속 작업의 정확도 향상에는 기여하지 못하고, 특히 분산 학습 환경에서는 계산 효율성만 저하된다는 것을 실험적으로 확인했다. 이는 접촉 지도 예측 전에 명시적인 대칭화가 적용되기 때문으로 보인다. 이후 pair_mlp_block에서는 ReLU 활성화 함수와 2배 채널 확장을 적용한 표준적인 2층 MLP를 통해 특성을 추가로 정제한다.

2.2.4 시퀀스 디코더 (Sequence Decoder)

시퀀스 디코더는 인코더의 구조를 반영하되, 반대 방향으로 작동한다. 즉, 128bp 해상도의 임베딩(트랜스포머 타워의 출력, 1536채널)을 점진적으로 업 샘플링하여 원래의 1bp 해상도로 복원하며, 동시에 특성 채널 수는 점진적으로 줄여서 768채널로 만든다. 이 작업은 총 7단계의 반복적 단계를 통해 수행되며, 각 단계마다 upres_block 모듈이 적용된다. U-Net 아키텍처에서 일반적으로 그렇듯이, 디코더는 인코더에서 동일한 해상도에서 저장된 중간 임베딩을 스킵 연결로 더하여 활용한다.

각 upres_block은 이전 단계의 출력과 인코더에서 같은 해상도에 해당하는 스킵 연결을 입력으로 받는다. 먼저, conv_block이 임베딩의 채널 차원을 스킵 연결의 채널 차원과 일치하도록 줄인다. 여기서도 마찬가지로, 채널 차원 감소를 처리하기 위해 잔차 연결이 크로핑 방식으로 적용된다. 그 결과는 시퀀스 길이 축을 따라 2배로 업샘플링 되는데, 이는 반복 방식으로 수행되며, 각 해상도 수준별로 학습 가능한 계수로 스케일링된다. 동시에, 인코더에서 들어오는 스킵 연결도 커널 너비 1을 가진 별도의 conv_block을 거쳐 처리된다. 처리된 스킵 연결은 이렇게 스케일링된 업 샘플링 결과에 더해진다. 마지막으로, 이 결합된 텐서는 다시 한 번 내부 잔차 연결이 포함된 conv_block을 통과한 후, 다음 단계로 전달되거나 최종 1bp 해상도 시퀀스 임베딩으로 반환된다.

2.2.5 출력 헤드 (Output Heads)

AlphaGenome 아키텍처는 입력된 DNA를 인코더, 트랜스포머, 디코더 단계를 거쳐 처리한 후, 여러 해상도에서 중요한 내부 표현을 생성한다. 이 학습된 임베딩(특히, 시퀀스 상의 1bp 및 128bp 해상도의 1차원 임베딩과, 2048bp 해상도의 2차원 쌍별 임베딩)은 다양한 유전체 특징을 예측하는 기반이 된다. 예측은 특정 생물종(인간 또는 생쥐)에 맞게 조정되며, 이를 위해 생물종별로 학습된 임베딩을 해당 함수 내에 통합하여 사용한다.

이들 공유 임베딩(embeddings_1bp, embeddings_128bp, embeddings_pair)을 생성할 때 적용되는 생물종별 임베딩 외에도, 각 출력 헤드는 인간과 생쥐 예측을 위해 별도의 독립적인 파라미터 세트를 학습한다. AlphaGenome의 전체 학습 손실은 각 헤드별로 정의된 손실의 합으로 구성되며, 추가적인 가중치 계수는 사용하지 않는다.

3 토의(Discussions)

AlphaGenome은 게놈의 조절 코드를 해독하려는 연구를 한층 더 발전시키며, 메가 베이스 규모의 DNA 서열로부터 다양한 기능적 유전체 신호를 동시에 예측할 수 있는 통합 딥러닝 모델이다. 이 모델은 각 분야의 특화된 SOTA 모델들과 비교해 동등하거나 더 뛰어난 성능을 보여주며, 이는 DNA에 암호화된 기본적인 조절 원리를 견고하게 학습하고 있음을 보여주는 동시에, 비부호화 변이의 기전적 해석에도 유용하다. AlphaGenome의 핵심 강점 중 하나는 효율적인 멀티모달 변이 효과 예측 능력으로, 단 한 번의 추론만으로 모든 예측 모달리티에 걸쳐 변이의 영향을 동시에 평가할 수 있다. 이와 같은 통합된 예측 능력은 복잡한 작용 메커니즘을 지닌 변이를 이해하는 데 필수적이며, TAL1 온코진 변이 효과의 정확한 재현 사례에서도 그 유용성이 잘 드러난다. 이러한 특성은 게놈 전반에 걸쳐 조절 서열 요소를 정밀하게 분석하는 대규모 연구에 강력한 기반이 될 수 있다. 또한, AlphaGenome은 스플라이스 접합부를 직접 모델링하는 새로운 기능을 갖추고 있어, 스플라이싱을 방해하는 변이에 대한 보다 전체적이고 포괄적인 이해를 가능하게 한다.

향후 AlphaGenome은 다양한 생물학 분야에서 폭넓게 활용될 것이다. 분자 생물학 분야에서는 AlphaGenome이 컴퓨터 기반 실험을 위한 엔진으로 활용되어, 가설을 빠르게 생성하고 wet-lab 실험의 우선순위를 정해줄 수 있다. 희귀질환 진단에서는, AlphaGenome의 향상된 변이 효과 예측 기능이 현재의 annotation 파이프라인에 추가적인 기능적 근거를 제공함으로써, 의미 불확실한 비부호화 변이에 대한 새로운 진단 경로를 찾아줄 수 있다. 또한, AlphaGenome의 스플라이싱, 유전자 발현, 크로마틴 접근성에 대한 향상된 예측은, 치료용 안티센스 올리고뉴클레오타이드 설계나 조직 특이적 인핸서 개발과 같은 서열 설계 응용에 “in-the-loop”으로 통합되어 활용될 것이다. 마지막으로, AlphaGenome은 DNA 서열을 기반으로 학습된 생성 모델과도 상호보완적으로 작용하여, 새롭게 생성된 서열의 기능적 특성을 예측하는 데 도움을 줄 것이다.

4 강력한 연구 도구

AlphaGenome의 예측 능력은 여러 연구 분야에서 활용될 수 있다:

① **질병 이해:** 유전적 교란을 더 정확히 예측함으로써, AlphaGenome은 질병의 잠재적 원인을 더 정밀하게 규명하는 데 도움을 줄 수 있으며, 특정 형질과 연관된 변이의 기능적 영향을 더 잘 해석함으로써 새로운 치료 타깃을 발견할 가능성도 열어준다. 특히, 희귀한 멘델형 질

환을 유발할 수 있는 드물지만 큰 영향을 미치는 변이 연구에 적합하다.
② **합성 생물학:** 이 모델의 예측 결과는 특정 조절 기능을 갖는 합성 DNA 설계에 활용될 수 있다. 예) 어떤 유전자가 신경세포에서만 활성화되고 근육세포에서는 비활성화되도록 설계.
③ **기초 생명과학 연구:** AlphaGenome은 유전체의 중요한 기능 요소를 매핑하고, 이들의 역할을 정의하는 데 도움을 줌으로써, 특정 세포 유형의 기능을 조절하는 데 가장 중요한 DNA 지시 요소를 식별하는 등 유전체에 대한 근본적인 이해를 가속화할 수 있다.

5 한계점

비록 AlphaGenome이 여러 면에서 진전을 이루었지만, 여전히 기존의 서열 기반 모델들과 공통된 과제를 안고 있으며, 모델 자체의 적용 범위에도 한계가 존재한다. 우선, 10만 염기쌍 이상 떨어진 장거리 조절 요소의 영향을 정확히 포착하는 것은 여전히 해결해야 할 과제이다. 또한 이 모델은 조직 및 세포 유형 특이적인 게놈 트랙을 어느 정도 예측할 수는 있지만, 세포 환경 전반에 걸친 조직 특이적 패턴을 정확하게 재현하거나, 상태-특이적 변이 효과를 예측하는 데에는 여전히 어려움이 있다. 현재 AlphaGenome은 인간과 생쥐에 국한되어 있으며, 본 논문에서의 모델 평가 또한 주로 인간을 중심으로 수행되었다. 또한, 개인 맞춤형 게놈 예측에 대해서는 아직 벤치마크를 수행하지 않았으며, 이는 이 분야의 모델들이 공통적으로 가지는 잘 알려진 한계점이다. 마지막으로, AlphaGenome은 변이의 분자적 결과를 예측하는 데 중점을 두기 때문에, 복합 형질 분석에는 적용이 제한적이다. 이러한 형질은 유전자 기능, 발생 과정, 환경 요인 등 서열 기반 기능 예측의 범위를 넘어서는 폭넓은 생물학적 과정을 포함하고 있기 때문이다.

이러한 한계들을 해결하는 것은 향후 연구 방향을 설정하는 데 중요하다. 주요한 연구 방향으로는 다음과 같은 것들이 있다: 변이 예측의 정확도와 활용도 향상(이를 위해 작업 특화 보정, 교란 실험 데이터셋에 대한 파인튜닝, 또는 단일세포 데이터 통합과 같은 접근 활용), 더 넓은 범위의 데이터 모달리티 통합(예를 들어, DNA 메틸화, DNA/RNA 구조적 특성과 같은 정보 추가), 모델의 근본적인 성능 개선(예컨대, DNA 언어 모델의 활용, 다종 확장, 그리고 실험 바이어스를 보정하기 위한 강건한 기법의 개발). 또한, AlphaGenome의 예측 결과를 보존 기반 점수, 유전자 기능 및 경로에 대한 기존 데이터와 통합하는 것은, 일반 변이와 희귀 변이 분석을 더욱 발전시키는 데 유용하다.

한마디로, AlphaGenome은 조절 게놈 분석을 위한 강력하고 통합된 모델로서, DNA로부터 분자 기능과 변이 효과를 예측하는 능력을 크게 향상시켰으며, 생물학적 발견과 생명공학 응용에 유용한 도구를 제공한다. 궁극적으로, AlphaGenome은 DNA 서열에 암호화된 복잡한 세포 과정을 해독하려는 보다 광범위한 과학적 목표를 향한 기초적인 발걸음이다.

AI 과학자의 세상 변혁을 위한 요건

How Far Are AI Scientists from Changing the World?

Q. Xie, et al., https://doi.org/10.48550/arXiv.2507.23276

요약: 대규모 언어 모델의 등장은 자동화된 과학 발견을 새로운 단계로 끌어올리고 있으며, 이제는 LLM 기반 AI 과학자 시스템이 과학 연구를 선도하고 있다. 머지않아 인간이 아직 발견하지 못한 새로운 현상을 규명할 수 있는 인간 수준의 AI 과학자가 현실이 될 것이다. 본 조사에서는 **"AI 과학자는 세계를 바꾸고 과학 연구 패러다임을 재편할 수 있을 만큼 얼마나 가까이 와 있는가?"**라는 핵심 질문에 초점을 맞춘다. 이를 위해, 현재의 AI 과학자 시스템이 달성한 성과를 종합적으로 분석하고, 주요 병목 지점과, 중대한 난제를 해결할 혁신적 발견을 만들어낼 수 있는 과학 에이전트의 등장에 필요한 핵심 구성 요소들을 찾고자한다. 이 조사가 현재 AI 과학자 시스템의 한계를 보다 명확히 이해하게 하고자 한다. 즉, 현재 어디에 서 있으며, 무엇이 부족한지, 그리고 과학 AI가 지향해야 할 궁극적 목표가 무엇인지 제시한다.

1 서론

과학적 발견은 인류 문명의 발전을 이끄는 근본적인 동력이다. 모든 주요 과학적 혁신은 인간 이해의 경계를 확장하고 사회적 진보를 촉진해 왔으며, 과학 연구의 주체인 인간은 체계적인 탐구 과정을 따라 왔다. 이 과정은 일반적으로 **(지식획득)** 관찰과 학습을 통한 과학적 지식 기초 마련 단계, **(아이디어 생성)** 해결되지 않은 질문을 다루기 위한 과학적 가설 수립 단계, **(검증 및 반증)** 강도 높은 실험과 이론적 분석을 통한 가설의 엄격한 검증 및 반증하는 과정을 거친다. 마지막으로, **(지속적 진화)** 실험 또는 이론적 결과에 기반하여 연구가 지속적으로 정교화되어, 과학 지식의 끊임없는 발전이 이루어진다.

연구 과정은 인간의 한계(제한된 시간과 인지 능력)에 의해 제약되며, 느린 문헌 조사, 좁은 지식 범위, 편향된 가설, 비효율적 실험으로 이어진다. 이 문제를 해결하기 위해 과학 발견의 자동화에 대한 관심이 높아지고 있다. 초기에는 딥러닝 기술을 활용하여 지식 획득을 지원하는 데 주력했다. 예) 사전학습 언어 모델에 도메인 특화 지식을 적용하여 과학 정보를 더 잘 표현하도록 하는 연구. 그러나 모델 용량의 제약과 고품질 과학 데이터의 부족으로 인해 이러한 방법들은 주로 과학적 도구로서 기능하며, 연구 과정의 일부만을 가속화하는 데 그쳤다. 최근 몇 년 동안 대규모·고품질 코퍼스로 학습된 LLM은 긴 텍스트 환경에서 뛰어난 이해 및 생성 능력을 보여주며, 축적된 지식을 바탕으로 새로운 과학적 가설을 창출하고, 자동화된 실험 도구를 활용해 이를 엄밀하게 평가하는 것이 가능해졌다. 이러한 진전은 자동화된 과학 발견을 새로운 단계로 끌어올리고 있으며, LLM 기반 AI 과학자 시스템이 과학 탐구의 전면에 나서게 하고 있다. AI 과학자가 인간의 한계를 넘어서는 능력(예: 제한된 기억 용량 극복)을 갖추게 되면, 의학, 에너지, 환경 등 다양한 분야의 중대한 난제를 해결하는 혁신적 발견을 이룰 잠재력을 가질 것이다. 이제 인류는 다음과 같은 질문을 탐구할 시점에 와 있다: **"AI 과학자는 세계를 변화시키고 과학 연구 패러다임을 재편하기까지 얼마나 가까이 와 있는가?"**

이에 답하기 위해, 기존 성과들에 대한 전망 중심(prospect-driven) 리뷰를 제공한다. Fig. 1은 성숙한 AI 과학자에게 필요한 능력의 계층 구조를 제시하며, 이는 현재와 미래의 관련 연구를 위한 타당한 로드맵이라고 볼 수 있다. 본 조사는 AI 과학자의 발전 단계를 체계적으로 정의하는 능력 기반 프레임워크를 네 가지 단계로 소개한다. **(1)지식 획득:** 기존 연구에서 과

학 지식을 검색, 검토, 이해하는 자율적 능력으로, AI 과학자의 기초적 역량이다. **(2) 아이디어 생성:** 혁신적이고 실현 가능한 가설을 대규모로 생성하는 능력이다. 이 능력은 AI 과학자 시스템을 기존의 자동화된 과학 도구와 구별하는 핵심 특징이다. **(3) 검증 및 반증:** AI가 생성한 과학적 가설을 시험하고 잠재적으로 반증하기 위해 실험을 체계적으로 설계·구현·분석하는 능력으로, AI 과학자를 단순한 아이디어 생성기에서 자율적 과학 지능으로 전환시키는 역할을 한다. **(4) 진화:** 내부 및 외부 피드백을 기반으로 전체 연구 능력을 지속적으로 향상시키는 능력으로, 초기 단계의 AI 과학자가 성숙한 과학적 에이전트로 발전하는 데 필수적이다. 여기서는 이 능력 프레임워크의 관점에서 현재 연구를 비판적으로 분석하며, 자율적 과학 지능이 혁신적 발견을 만들어내기 위해 필요한 핵심 병목 지점과 부족한 요소들을 식별한다.

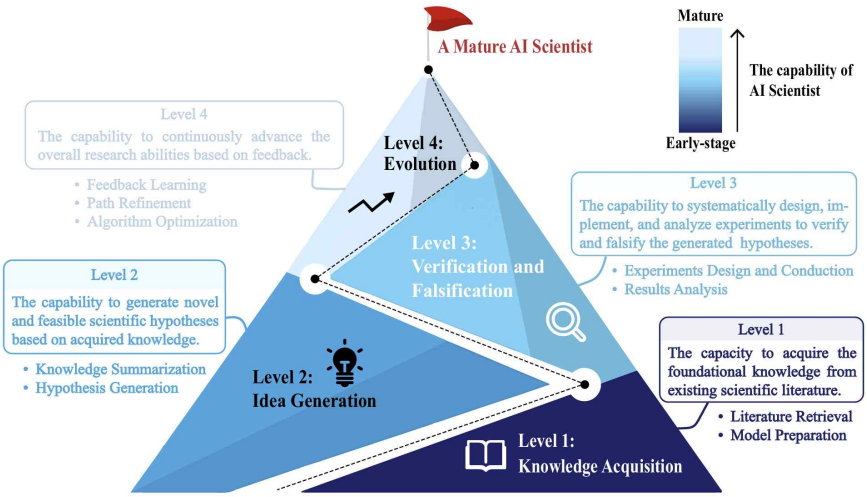


Figure 1: The capability level of an AI Scientist, illustrating the progression from foundational knowledge acquisition (Level 1), through idea generation (Level 2), rigorous hypothesis verification and falsification (Level 3), to continuous evolution (Level 4). We outline the core functions for each capability level.

앞서 언급한 능력들과 관련하여, **다양한 시도와 성과, 그리고 한계점들이 존재한다(Fig. 2)**. 여기서는 현재 AI 과학자 시스템이 갖춘 **지식 획득(2장)**, **아이디어 생성(3장)**, **검증 및 반증(4장)** 능력에 대한 기존 성과를 체계적으로 검토하며, 대표적인 방법들을 각 능력 수준에 대응시켜 정리한다. 이어 **AI 리뷰어 시스템**의 발전을 개괄하고, 현재의 AI 과학자 시스템이 독립적으로 고품질의 과학적 발견을 수행할 **능력이 부족함을 보여주는 실증적 근거**를 제시한다**(5장)**. 이러한 한계를 극복하기 위해서는 AI 과학자 시스템이 **진화 능력(6장)**을 갖추는 것이 필수적이며, 이는 과학 연구 패러다임을 재편하는 방향으로 나아가기 위한 핵심 요소이다. 또한 현재 AI 과학자 연구가 직면한 **도전 과제들(7장)**을 포괄적으로 살펴보고, **향후 연구 방향과 질문들(8장)**을 논의함으로써, 진정한 자율적 과학 지능을 향한 합리적인 로드맵을 제시한다.

2 지식 획득

정의: 지식 획득은 AI 과학자의 기초적 능력으로, 기존 과학 문헌에서 분야별 지식을 스스로 검색·검토·이해하여 연구를 위한 과학적 지식 기반을 체계적으로 구축하는 자율적 능력.

지식 획득은 AI 과학자의 근본적인 능력으로, 기존 과학 연구로부터 지식을 스스로 검색/검토/이해하는 자율적 능력을 포함한다. 이후의 연구 활동에 필요한 지식 기반을 구축하는 역할을 하며, 인간 연구자의 문헌 조사와 유사하다. 현대 AI 과학자 시스템(LLM 기반 방법들)이 다음

의 두 가지 도전을 어떻게 해결하고 있는지 체계적으로 살펴본다: (1) 문헌 검색/선별: 방대한 코퍼스에서 관련 연구를 필터링하는 문제, (2) 지식 추출/요약: 구조화된 통찰을 도출하는 문제. LLM 등장 이전에도, AI 과학자 시스템의 기본 개념은 이미 상당한 관심을 받아왔으며, 많은 소규모 언어 모델들이 지식 획득 단계에서 주목할 만한 성과를 보여왔다. 따라서 본 절에서는 AI 과학자 시스템의 지식 획득 연구를 LLM 이전 시대(대략 2020년 이전)와 LLM 시대(2020년 이후, GPT-3와 같은 모델 등장 이후)로 구분하여 정리한다.

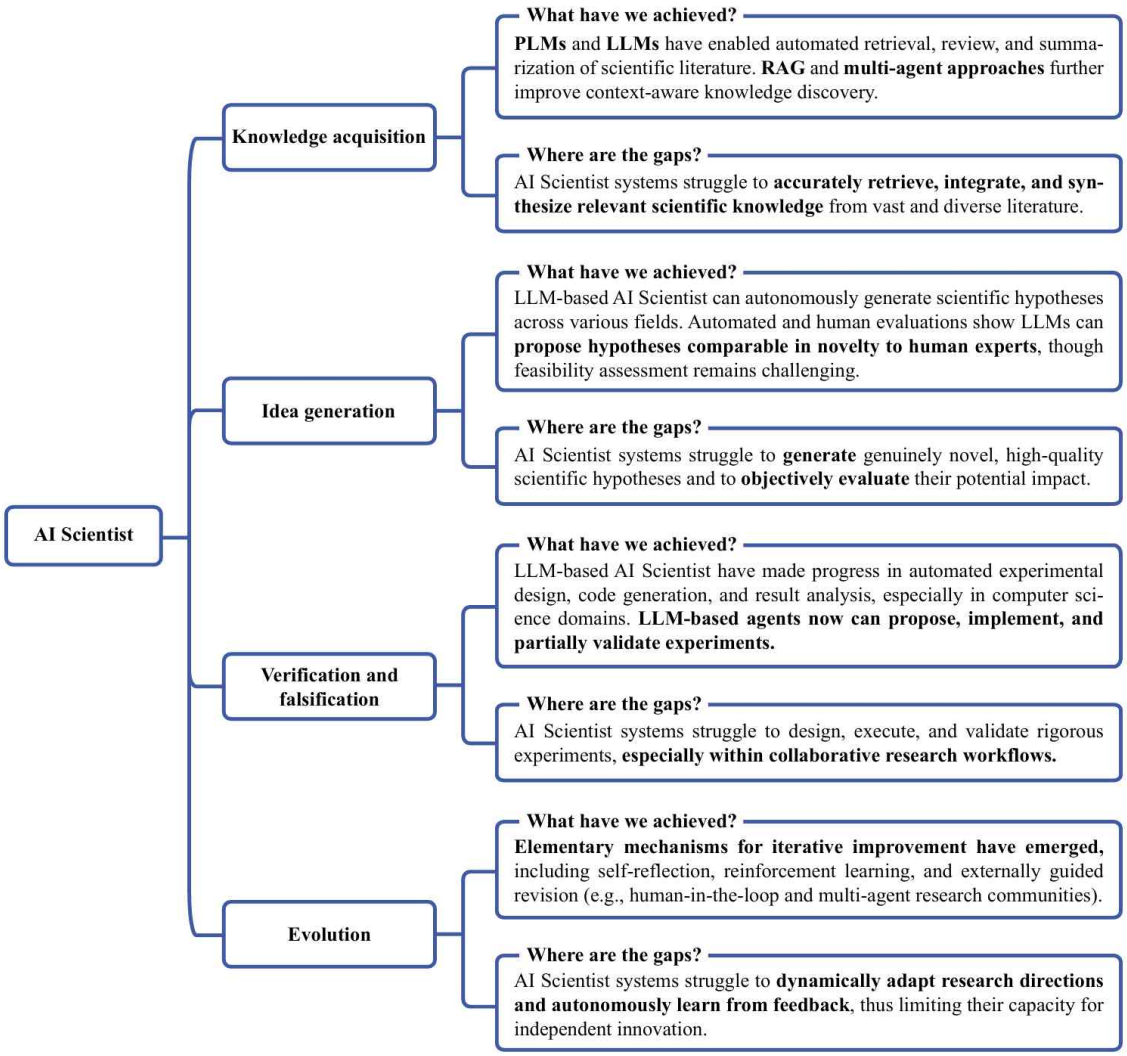


Figure 2: The current capability landscape of AI Scientist systems across four progressive levels. We summarize the current achievements for each level and highlight critical gaps before AI Scientist systems can autonomously make ground-breaking scientific discoveries.

2.1 LLM 이전 시대 방법들

연구자들이 소규모 PLMs(pre-trained language models)을 분야 특화 지식으로 미세 조정하여 상당한 수준의 과학적 표현을 학습시키는 방식으로 자동화된 지식 획득에서 큰 진전을 이루었다. 이 모델들은 과학 문헌으로부터 구조화된 텍스트 마이닝과 지식 추출에서 뛰어났으며, 이후 LLM 기반의 더욱 발전된 방법들이 등장할 수 있는 기반을 마련했다.

문헌. LLM이 등장하기 전, 초기 자동화된 지식 획득 시스템은 주로 대규모 과학 코퍼스로 학습된 소규모 PLM에 의존했다. SciBERT: 114만 편의 논문으로 사전학습된 최초의 PLM으로, 과학 분야 NER 및 인용 분류 작업에서 강력한 성능을 보이며 특히 컴퓨터 과학 분야에서 연구자의 연구 효율을 크게 향상. SCHOLARBERT: 221B 토큰의 학술 텍스트로 사전학습되어 과학 문헌의 문맥 이해력을 크게 강화. ClinicalBERT, PubMedBERT, BioELMo, BioBERT, BioMegatron, BioM-Transformers, MatSciBERT와 같은 분야 특화 모델들: 고품질 임상·생물학·화학 코퍼스로 사전학습. 이러한 모델들은 구조화된 텍스트 마이닝에 집중하여, 규칙 기반 파싱이나 그래프 기반 인용 분석과 같은 기술을 활용해 과학 논문의 초록이나 본문에서 핵심 구절, 개체, 관계 등을 추출한다. Kang 등 (2022)은 저자 네트워크 그래프 기법을 활용하여 사용자의 논문 상호작용 뒤에 숨겨진 깊은 연관성을 밝혔다. Kang 등 (2023)은 사용자 선택 논문을 기반으로 저자 위원회를 구성하고, 여러 출처의 신호를 수용하여 관련 문헌을 동적으로 탐색하는 시스템을 구축한다. 위와 같은 LLM 이전 시대의 연구들은 자동화된 지식 획득에서 주목할 만한 성과를 달성했으나, 확장성과 문맥 이해 능력이 본질적으로 제한되어 있었으며, 이는 이후 더욱 정교한 지식 획득 접근 방식이 등장하는 기반을 마련했다.

실험. 연구자들이 실험 결과(예: 리더보드 구축)에서 지식을 추출하는 방법도 탐구했다. NLP-progress나 Papers with Code와 같은 데이터 소스를 활용하려는 노력에 이어, SciGen과 Numerical-NLG는 표 설명 생성에서 언어 모델의 능력을 평가하는 벤치마크이다. 그러나 이러한 연구들은 엄격한 품질 보증 체계가 부족한 것으로 드러났다. 예) 여러 리더보드 간의 과학적 개체 표준화가 이루어지지 않았으며, 관련 출판물의 커버리지 또한 불완전했다. 반면, SCICM은 다양한 분야의 약 240만 개 학술 논문을 포함하는 공개 접근 arXiv를 선택하여, 더 많은 관련 출판물을 확보한다. SCICM과 유사하게, TDMS-IE와 AxCell도 실험 결과 정보를 추출하여 “Task-Dataset-Model” 삼중항(triple)과 실험 결과 개체를 활용해 자동으로 리더보드를 구축한다. 이러한 연구 흐름은 TELIN과 ORKG-Leaderboards 같은 후속 연구들로 확장되었으며, 연구 효율성을 향상시키려는 뚜렷한 추세를 보여준다.

2.2 LLM 시대 방법들

LLM은 AI 과학자 시스템의 지식 획득 능력을 획기적으로 변화시켰다. 이러한 진전은 문헌 검색 도구, 아이디어·실험 중심의 지식 추출 기법, 특화된 평가 벤치마크에 이르기까지 광범위하나, 검색 정밀도와 다중 문서 요약에서 여전히 지속되는 도전 과제들을 함께 드러낸다.

2.2.1 문헌 검색과 선별

연구자들은 다양한 출처에서 관련 과학 문헌을 효율적으로 필터링, 선택, 추천할 수 있는 특화된 툴킷과 시스템을 개발해 왔다. 이들 연구 중, PaperWeaver와 같은 시스템은 문헌 검색 파이프라인 전반의 여러 하위 단계를 지원할 수 있다. 이러한 진전을 바탕으로, DORA AI Scientist는 핵심 검색 과정에 RAG를 통합하여 문맥적 관련성과 자동화를 향상시켰다. 그럼에도 불구하고, 모든 시스템이 완전한 자율성을 갖춘 것은 아니다. 예) CodeScientist와 같은 반자동 프레임워크는 여전히 아이디어 생성 단계에서 사용자가 제공한 논문 목록에 의존한다. 즉, 진정한 자동화 및 고정밀 문헌 발견을 달성하는 데 여전히 과제가 남아 있다.

이러한 접근 방식을 기반으로, 연구자들은 과학 문헌을 검색하고 필터링하기 위한 다양한 구현 방안을 개발해왔다. Schmidgall 등은 arXiv API를 활용하여 논문을 검색하고, 요약, 전문 추출, 선별 작업을 수행한다. 한편, Jiabin 등은 arXiv에서 인용 횟수가 많은 분야별 논문을

대상으로 인용 기반 필터링을 적용한다. 단일 에이전트 기반 검색을 넘어, 다중 에이전트 시스템도 탐구되고 있다. Gottweis 등은 웹 검색 도구를 통합하고, Lu 등은 Semantic Scholar의 API를 활용하여 가설 생성에 적합한 관련 논문을 식별한다. RAG 분야에서는 Lála 등과 Skarlinski 등이 PaperQA와 PaperQA2를 과학적 질문 응답 과제에 도입하였다. 또한, PaSa는 세션 수준 강화 학습을 적용한 이중 에이전트 구조를 통해 복잡한 쿼리에 대해 논문을 자율적으로 검색·분석하고 인용 네트워크를 탐색한다. 그럼에도 불구하고, Beel 등은 기존 시스템이 여전히 키워드 기반 매칭에 크게 의존하며, 종종 고전적이지만 잠재적으로 최신이 아닌 논문을 검색하는 문제가 있음을 지적하며, 문맥을 고려한 검색 전략의 필요성을 강조한다.

2.2.2 지식 검색 및 요약

잘 정제된 문헌을 기반으로, 지식 검색 및 요약 방법은 구조화된 데이터와 비구조화된 데이터 모두에서 지식을 도출하는 것을 목표로 한다. 이 연구 분야는 주로 두 가지로 나뉜다: **(1) 아이디어 지향 접근법:** 연구 논문에서 핵심 개념과 통찰을 추출한다. **(2) 실험 지향 접근법:** 실험 결과를 검색하고 이를 종합적인 보고서로 요약한다. 이 단계에서 LLM은 문헌에서 가치 있는 지식을 처리하고 체계화하는 데 중추적 역할을 수행한다.

아이디어 지향 접근법은 LLM을 활용하여 과학 문헌에서 핵심 개념과 통찰을 추출한다. 이들을 다음으로 분류하였다. **(1) 직접 프롬프팅**은 기본적인 접근법으로, LLM 기반 작업 수행이 명시적인 프롬프트에 의해 직접적으로 안내된다. 그러나 컨텍스트 길이의 제약이나 검색 정확도 문제 등으로 인해 실제로는 거의 사용되지 않으며, 보통 baseline 방법으로 활용된다. **(2) 문헌 구조 기반 방법**은 과학 문헌의 고유한 구성이나 논문의 구조(예: 초록-서론-결론(AIC))에 기반한 사전 정의된 계획을 활용한다. **(3) RAG 기반 방법**은 관련 외부 정보를 먼저 검색한 뒤 이를 생성의 컨텍스트로 사용함으로써 LLM의 요약 능력을 강화하여 더 사실적이고 시의성 있는 출력을 제공하는 것을 목표로 한다. 예) LitLLM은 문헌 검색과 지식 추출 모두에 RAG를 활용. **(4) 멀티 에이전트 기반 방법**은 여러 에이전트 시스템의 협업을 통해 문헌 리뷰를 생성한다. 예) Schmidgall/Moor는 지식 검색/요약 가속화를 위해 여러 LLM을 통합.

실험 지향 접근법은 과학 문헌에서 실험 결과를 추출하여 이를 종합적인 보고서 형태로 요약하는 절차를 진행한다. LLM은 구조화된 데이터를 분석하고, 실험 설계를 해석하며, 포괄적인 요약을 생성하는 데 핵심적인 역할을 한다. 그러나 이 작업은 Task-Model-Dataset 삼중 구조의 정확한 추출이 필요하고, 연구 벤치마크가 지속적으로 변화한다는 점에서 복잡하다. 예) LEGO-bench: SOTA 연구를 지속적으로 반영하려는 과학 리더보드 생성 시스템을 평가하는 벤치마크를 제안. Park 등: 여러 데이터 셋에서 실험 데이터를 반자동으로 추출하여, 비교 기반 접근이 새로운 과학적 발견을 이끌어냄. 한편, Wu 등은 LAG(Leader Auto Generation)를 제안하며, 표 검색, 데이터 통합, 리더보드 생성, 그리고 가장 효과적인 리더보드를 선별하기 위한 품질 기반 선택까지 포함하는 포괄적인 4단계 동적 리더보드 생성 프로세스를 구현.

2.3 평가

평가 벤치마크는 AI 과학자 시스템의 지식 획득 능력을 평가하는 데 중요한 역할을 한다. 이러한 벤치마크는 정보검색 정확도, 요약품질, 지식구조 생성 등 다양한 측면에 초점을 맞춘다.

고전적인 과학 논문 요약: 지식 획득 능력의 초기 평가는 문헌 요약 연구에서 비롯되었다. SCITLDR: 과학 논문에 대한 5,400개 이상의 요약을 포함하며, 모델이 매우 압축된 형태의

과학 논문 요약을 생성하도록 요구. X-SCITLDR: 다국어 데이터셋으로, 서로 다른 교차언어 요약 전략의 효과성을 평가한다. 이러한 벤치마크는 원본 과학 논문의 일부만을 테스트 항목으로 활용하며, 상대적으로 짧은 컨텍스트 윈도우를 가진 모델을 목표로 한다.

문헌 검색: 지식 획득 평가의 또 다른 중요한 측면은 모델이 관련 과학 문헌을 검색하는 능력이다. CiteME: 최근 제안된 벤치마크로, 주어진 문맥을 바탕으로 올바른 논문을 인용할 수 있는지를 평가하여 LLM 기반 과학 에이전트를 측정한다. LitSearch: 과학 문헌 검색 시스템의 능력을 평가하는 또 다른 벤치마크로, 597개의 수작업으로 정제된 질의와 6만4천개 이상의 논문을 포함하여, 복잡한 질의 이해 및 관련 문서 탐색 능력을 평가한다.

자동화된 문헌 리뷰 및 도구: 최근의 벤치마크들은 LLM이 자동 문헌 리뷰 수행과 과학적 지식 조직능력을 평가하기 시작했다. 예) Yun 등은 문헌 리뷰 작업에서 연구자들이 블랙박스 시스템보다 투명한 AI 도구를 선호한다는 사실을 발견했다. 이를 바탕으로 Hsu 등은 CHIME을 제안했는데, 이는 주제(category) 생성 및 연결 능력을 평가하는 반자동 생성 벤치마크이지만, 여전히 LLM들은 구체적인 연구 결과(논문)를 적절한 주제에 정확히 배정하는 데 어려움을 겪는다. 유사하게, Agarwal 등은 arXiv 논문에서 파생된 새로운 평가 프로토콜과 테스트 세트를 제안하여, LLM이 문헌 리뷰를 실제로 작성할 수 있는지를 엄격하게 평가한다. 특히 제로샷 무결성 유지와 모델이 발전함에 따라 발생하는 테스트셋 오염 방지에 중점을 둔다.

3 아이디어 생성

정의: AI 과학자의 아이디어 생성 능력이란, 획득한 지식과 기존 방법을 바탕으로 혁신적이고 실현 가능한 과학적 가설을 자율적으로 수립하는 능력을 의미한다. 이 능력은 AI 과학자 시스템을 단순한 자동화된 과학 도구와 구별하는 핵심 요소이다.

전통적 연구 패러다임에서는, 연구자가 아이디어를 제안하고 AI는 아이디어 구현이나 정제와 같은 노동집약적 작업을 수행한다. 즉, AI는 과학 연구의 주체가 아니라 자동화된 도구로 간주된다. 그러나 LLM의 등장으로, AI는 이제 방대한 과학 문헌으로부터 도메인 특화 지식을 자율적으로 검색·검토·이해할 수 있게 되었다(2절). 그 결과, 혁신적이고 실행 가능한 가설을 자율적으로 생성할 수 있는 AI 과학자 시스템을 개발하려 하고 있다. 이 새로운 패러다임에서는, 자동화된 과학 도구와 달리 AI 과학자 시스템이 과학 연구의 핵심 추진 주체가 되어 새로운 가설을 제시하고, 인간이나 자동화된 실험 도구는 이러한 가설을 실행하는 역할을 맡는다.

3.1 방법

아이디어 생성에 대한 고전적 형식화는 Swanson(1986)의 “ABC모델”로 거슬러 올라간다. 이 모델에서 A와 C라는 두 개념은 문헌 속에서 어떤 중간 개념 B와 모두 동시에 등장한다면 서로 연결되어 있을 것이라고 가정한다. 이 연구에 영감을 받아, 이후 연구들은 그래프 모델이나 단어 벡터를 활용해 개념 쌍 간의 연결을 식별함으로써 과학적 가설을 생성하고자 했다. 그러나 복잡한 과학적 아이디어를 이진적 형태의 개념 연결로 환원하는 방식은 입력의 유형을 제한할 뿐 아니라, 생성된 가설의 표현력을 크게 제한하는 문제가 있다.

LLM의 급속한 발전으로 인해 연구자들은 특수한 프롬프트 전략, 포스트 트레이닝 기법, 또는 멀티 에이전트 협력 방식을 활용해 LLM이 자연어 형태로 혁신적인 과학적 가설을 생성하는 가능성을 탐구하고 있다. 예) Li 등: 사람이 연구를 수행하는 방식을 모방하여

COI(Chain-of-Ideas) 에이전트를 제안했다. 이 LLM 기반 에이전트는 관련 문헌을 체인 구조로 조직하여 연구 분야의 점진적 발전을 효과적으로 반영하고, 기존 문헌에 기반한 고품질의 가설을 생성하도록 한다. Si 등: RAG와 최근의 추론 시간 스케일링 기법을 결합한 LLM 에이전트를 활용한다. 이들은 LLM에 연구 주제마다 4,000개의 초기 아이디어를 생성하게 하고, 이후 reranker를 사용해 가장 고품질의 가설을 선택한다. 또 다른 주목할 만한 연구 흐름은 포스트 트레이닝 최적화에 집중한다. 예) Wang 등: 과학 문헌에서 동적으로 검색된 배경 정보를 기반으로 자연어 가설을 생성하는 SCIMON 프레임워크를 제안했다. SCIMON: T5 모델을 in-context 대조 학습 목표와 언어 모델링 목표로 파인튜닝하고, 이후 반복적 최적화 과정을 통해 생성된 가설의 새로움을 명시적으로 강화한다. 한편 AI Scientist(Lu et al., 2024)는 여러 차례의 chain-of-thought 추론과 self-reflection을 통해 각각의 가설을 정교하게 다듬고 발전시킨다. 개별 에이전트를 넘어 협력적 멀티 에이전트 시스템도 가능성을 보여주고 있다. 예) Yang 등: MOOSE-CHEM이라는 LLM 기반 멀티 에이전트 프레임워크를 개발(다음의 세 단계 수행)했다: (1) 화학 분야 문헌에서 '영감을 주는 논문'을 검색, (2) 이 영감을 활용해 목표 연구 질문과 관련된 가설 생성, (3) 생성된 가설을 식별하고 품질 기준에 따라 순위화.

3.2 평가

생성된 가설의 평가는 일반적으로 신규성(novelty), 실현 가능성(feasibility), 명확성(clarity), 흥미도(excitement) 등을 포함한 여러 기준에 따라 수행된다. 각 기준의 세부 내용은 표 1에 제시되어 있다. 이러한 평가 기준을 바탕으로, 연구자들은 일반적으로 AI가 생성한 가설의 품질을 추정하기 위해 두 가지 접근법-자동 평가와 인간 평가-을 사용한다.

자동 평가 방법은 보통 LLM을 자동 평가자로 활용하여 가설의 품질을 판단한다. 예) Chai 등은 Claude 3.5 모델을 사용하여 각 가설을 배경 문맥/연구 주제와 비교하고, 충분한 신규성, 과학적 타당성, 명확성을 갖추었는지 확인한다. Li 등은 Idea Arena라는 pairwise 평가 시스템을 제안했는데, 여기서 LLM 판정자가 가설을 쌍대 비교하여 순위를 매기고, 각 가설 생성 모델에 대한 ELO 점수를 계산한다. LLM을 평가자로 활용할 때의 성능은 학습 데이터의 다양성, 내재적 모델 편향, 평가 불확실성 등 다양한 요인에 의해 영향을 받는다.

Table 1: Commonly-used evaluation criteria for AI-generated scientific hypotheses.

Criterion	Description
Novelty	Whether the hypothesis is creative, distinct from existing work on the topic, and offers fresh insights.
Feasibility	How practical and achievable the hypothesis is as the basis for a research project.
Clarity	Whether the hypothesis is clearly stated and easy to understand.
Excitement	The potential excitement and impact the hypothesis could generate if pursued as a full research endeavor.

인간 평가는 가설을 평가하는 gold standard이다(정량적 지표만으로는 평가가 충분하지 않은 경우가 많기 때문). 예) Hu 등은 10명의 전문가 패널을 모집하여, 가설의 신규성과 전체적 품질을 기준으로 평가하게 했다. 또한, Si 등은 대규모 인간 연구를 수행했는데, 79명의 전문가 연구자를 모집하여 세 가지 조건(전문가 작성 아이디어, AI 생성 아이디어, 인간 전문가가 재순위화한 AI 생성 아이디어) 각각에서 49개의 아이디어를 블라인드 리뷰하도록 했다. 그 결

과, LLM이 생성한 아이디어가 인간 전문가가 작성한 아이디어보다 신규성 에서 더 높게 평가되었지만, 실현 가능성에서는 다소 낮게 평가되었다. 일부 연구에서는 참조기반 텍스트 생성 지표를 활용해 가설의 품질을 평가하기도 한다. 예) Qi 등은 BLEU와 ROUGE 점수를 사용하여 생성된 출력과 정답 참조 간의 단어 겹침 정도를 측정했다. Yang 등은 Matched Score를 채택하여, 6점 Likert 척도를 활용해 생성된 가설과 원래 가설 간의 유사도를 계산했다.

특히, 현재 대부분의 AI 생성 가설의 실현 가능성 평가 방법이 여전히 인간의 직관과 텍스트 설명에 기반한 주관적 판단에 크게 의존하고 있다. 이러한 방법은 명백히 실행 불가능한 가설을 효율적으로 걸러낼 수 있지만, 정확성과 객관성이 부족하다. 따라서 가설의 실현 가능성을 평가하기 위해서는 실험적 검증과 반증이 가능한 실질적 방법(예: 가설을 실행 가능한 코드로 변환)이 필요하다. 본 조사에서는, 새로운 과학적 가설을 검증하기 위한 실험 설계 능력을 보다 발전된 형태의 검증 및 반증 능력으로 분류하며, 이에 대해서는 4절에서 상세히 논의한다.

4. 검증과 반증

정의: 이 능력은 AI가 생성한 과학적 가설을 AI 과학자가 체계적으로 검증하고 반증하기 위해 실험을 설계, 실행, 분석할 수 있게 한다. 특히, 연구 사이클을 완성시키는 핵심 요소로, AI 과학자를 단순한 아이디어 생성자에서 자율적 과학 지능으로 변화시키는 역할을 한다.

현재 AI 생성 가설에 대한 평가는 특히 실현 가능성 측면에서 주관적 인간 평가나 피상적인 텍스트 유사도 지표에 크게 의존하고 있다. 이 접근법은 실질적인 실험적 검증이 제공할 수 있는 경험적 엄밀성을 본질적으로 결여하고 있으며, 이는 검증/반증 능력의 필수적 역할을 강조한다. 가설을 적극적으로 테스트하고 검증하는 능력(verification)과, 엄밀하게 도전하고 반증하는 능력(falsification)을 시뮬레이션 또는 실제 환경에서 수행할 수 있는 능력은, 폐쇄 루프(closed-loop) 과학적 발견을 수행할 수 있는 완전한 AI 과학자를 정의하는 핵심 요소이다.

전통적으로 과학적 검증은 인간 과학자가 실험을 설계, 실행, 분석하는 과정을 필요로 한다. 최근 들어 LLM과 자동화된 실험 도구의 급속한 발전으로, AI 과학자 시스템이 과학적 가설을 검증하고 반증하는 능력도 크게 향상되었다. 초기에는 직접 코드 실행과 결과 매칭과 같은 단순한 작업 중심의 접근법에 머물렀으나, 점차 더 정교하고 포괄적인 과정으로 발전하고 있다. 예를 들어, AI 과학자는 자동 코드 생성과 오류 수정 기능을 통해 실험을 설계하고 실행하며, 이를 통해 완전 자동화된 개방형(open-ended) 과학적 발견이 가능하다.

Type	Total citations	Avg. citations
Without Implementation	216	10.3
With Implementation	**325**	**25.0**

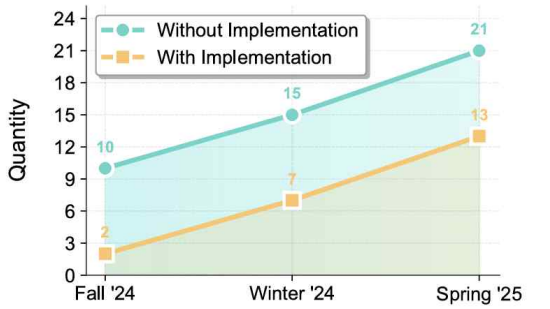


Figure 3: An analysis of the number of publications in the field of AI Scientist systems on arXiv. The upper panel displays the average number of citations up to now, categorized by containing implementation details. The lower panel shows the growth in the total number of these papers with the same categorization.

그러나 검증 및 반증 과정은 여전히 현재 AI 과학자 시스템에 있어 중요한 과제이다. 2025년 5월 23일까지 arXiv에 게재된 AI 과학자 관련 논문을 분석한 결과를 그림 3에 보였다. 이 분야의 출판물 수는 증가하고 있으며, 아이디어 생성에 초점을 맞춘 연구가 구현을 포함

한 연구보다 많지만, 실질적인 구현 세부 사항을 포함한 논문은 평균 인용 횟수가 훨씬 높다. 이는 실행 가능한 검증이 AI 과학자 커뮤니티에서 높은 평가를 받고 있음을 보여준다.

4.1 방법

검증과 반증의 궁극적인 목표는 AI가 생성한 가설이 실현 가능함을 평가하기 위해 실험을 설계하고, 구현하며, 분석하는 것에 주로 기반한다. 그러나 화학, 생물학, 의학과 같은 과학 분야에서 실험 장소, 장비, 실험 안전성과 같은 제약조건을 고려하면, 현재의 AI 과학자 시스템은 컴퓨터 과학, 특히 머신러닝 분야에서 아이디어 검증과 반증에 주로 초점을 맞추고 있다.

실험 설계 측면에서, AI 과학자 분야에서는 여러 혁신적인 접근이 등장하고 있다. BioPlanner는 생물학 분야에서 실험을 계획하는 LLM의 능력을 평가하기 위한 자동화된 평가 프레임워크를 도입했다. DSBench와 ScienceAgentBench는 데이터 기반 과학적 발견을 위한 실험 설계에서 에이전트의 역량을 평가한다. Yang 등 (2023)은 LLM을 “마스터브레인 과학자”(주연구자에 해당)로 활용하여, 중심적 역할을 맡아 완전 폐쇄형 연구를 이끄는 최초의 연구다. 이후의 AI 과학자 시스템들, 예를 들어 AI Scientist(Lu 등, 2024)는 LLM 에이전트를 활용해 연구 아이디어를 자율적으로 제안하고, 검증 가능한 가설을 수립하며, 고수준의 실험 계획을 설계한다. 이러한 계획은 일반적으로 변수를 정의하고, 데이터셋을 선택하며, 기준 모델을 고르고, 평가 지표를 설계하는 과정을 포함한다. 그러나 과학적 실험 설계의 다양성과 복잡성은 현재의 AI 과학자 시스템에 상당한 도전 과제를 제시한다. 이들 시스템이 생성하는 실험 설계는 종종 과학적 엄밀성, 혁신성, 실용성이 부족하여 고수준 연구의 요구를 충족하기 어렵다.

실험에서 이론적으로 가장 중요하고 어려운 부분인 구현은 종종 저장소(repository) 단위의 코드 생성 방식으로 다루어진다. 예) ToolGen과 같은 LLM 기반 기법은 자동 완성 도구를 통합하여 신뢰할 수 있는 종속성을 갖춘 저장소 단위의 코드를 생성한다. CoCoGen은 코드 저장소에서 추출한 정보를 활용하여 반복적으로 정렬을 수행하고 오류를 수정한다. CatCoder는 관련 코드와 타입 정보를 통합함으로써 저장소 수준 코드 생성을 향상시킨다. 에이전트 기반 방법은 더욱 강력한 성능을 보여주고 있다. CodeAgent는 다섯 가지 프로그래밍 도구를 통합하여, 정보 검색, 코드 심볼 탐색, 코드 테스트 등 소프트웨어 아티팩트와의 상호작용을 가능하게 한다. RepoCoder는 유사도 기반 검색기와 사전 훈련된 코드 언어 모델을 반복적인 검색-생성 파이프라인에 통합하여 저장소 수준 코드 자동 완성 과정을 단순화한다. 실제 시나리오에 더 가까운 RepoGraph는 라인(line) 단위에서 작동하여 기존의 파일 단위 브라우징 방법보다 더 세밀한 접근을 제공한다. 그래프의 각 노드는 코드 한 줄을 나타내고, 엣지는 코드 정의와 참조 간의 의존성을 나타낸다. RepoGraph는 SWE-bench에서 기존 방법 대비 평균 상대 개선률 32.8%를 달성하며 성공률을 높였다. SciCode는 과학적 저장소 수준 코드 생성을 여러 하위 문제로 분해하며, 각 문제는 지식 회상, 추론, 코드 합성을 포함한다. AIDE는 머신러닝 연구를 코드 최적화 문제로 형식화하고, 시행착오 과정을 잠재적 해 탐색 공간에서의 트리 탐색으로 모델링한다. MLR-Copilot은 실험 계획을 실행 가능한 실험 코드로 번역하는 ExperimentAgent를 제안하며, 실행의 복잡성을 관리하기 위해 인간 피드백과 반복적 디버깅 메커니즘을 필수적으로 통합한다. 더 높은 수준의 자율성을 추구하기 위해, LLM 기반 멀티 에이전트 프레임워크인 AutoP2C는 연구 논문의 텍스트 및 시각 정보를 처리하여 자동으로 실행 가능한 코드 저장소를 생성한다. 이러한 지속적인 발전은 과학적 구현의 본질적 어려움을 극복하기 위해 점점 더 정교한 방법들을 개발하려는 연구 흐름을 잘 보여준다.

분석은 실험 과정의 마지막 단계로서, **실험 결과를 잘 구성된 실험 보고서로 체계화**하는 데 핵심적인 역할을 한다. 최근에는 MCX-LLM과 같은 연구가 자연어 설명을 몬테카를로 시뮬레이션 실행을 위한 기계 판독 가능 입력으로 변환하는 LLM 활용 방식을 탐구하여 분석 효율을 향상시켰다. 이러한 시스템의 역량을 평가하기 위해 FigureQA, ArxivQA, MMSCI와 같은 벤치마크가 개발되었으며, 이는 그래프, 차트, 표 등 복잡한 시각 자료를 이해하고 추론하는 에이전트의 능력에 초점을 맞춘다. 데이터 해석뿐 아니라 LLM-ref 같은 도구는 여러 출처 문서의 정보를 통합하고 참고문헌 관리를 강화함으로써 연구자들이 논문 작성 과정에서 더 효율적으로 작업할 수 있도록 돕는다. 이러한 발전을 바탕으로 최신 AI 과학자 시스템들은 LLM 기반 코드 생성을 활용하여 견고한 통계 분석을 자동화하고 플롯, 표, 차트 등의 시각화를 생성한다. 이러한 자동화는 실험 결과 보고를 크게 간소화한다. 또한 여러 독립적인 실험 실행을 수행하고 결과를 집계하여 종합적인 메타 분석을 가능하게 한다. 분석 단계 이후에는 AI 과학자 시스템이 결과를 논문 형태로 통합하여, 방법론적 세부 사항, 데이터 분석, 결과, 결론을 포함한 표준 학술 형식의 구조화된 보고서를 자동으로 생성할 수도 있다. 자동 논문 작성은 이러한 시스템의 중요한 활용 사례이지만, 논문 생성 자체가 AI 과학자 시스템의 핵심 역량은 아니기에, 본 조사에서는 연구논문의 작성보다는 분석 및 추론 능력에 초점을 맞춘다.

4.2. 평가

실제로 강건한 검증과 반증을 달성하는 것은 현재의 AI 과학자 시스템에서 여전히 큰 격차로 남아 있으며, 이는 다양한 고난도 벤치마크에서 최신 LLM조차도 보이는 성능 격차를 통해 극명하게 드러난다. 이러한 벤치마크 작업에는 엔드 투 엔드 AI 모델 학습 워크플로 실행, 학술 논문의 방법 및 실험 결과 재현, 연구 논문에 포함된 복잡한 알고리즘 설명으로부터 실행 가능한 코드를 정확히 생성하는 작업 등이 포함된다. 예) MLE-Bench는 Kaggle 머신러닝 과제를 해결하는 능력을 평가하는데, OpenAI의 “o1-preview” 모델은 메달 수준 성공률 16.90%에 그쳤다. PaperBench는 ICML 연구 논문을 처음부터 재현하는 능력을 요구하며, OpenAI의 “o1-high”는 26.00% 재현 점수를 기록했다. 추가적인 어려움은 SciReplicate-Bench에서도 확인되는데, 이는 알고리즘 설명으로부터 실행 가능한 코드를 생성하는 능력을 평가하며 Claude-Sonnet-3.7은 39.00% 실행 정확도에 그쳤다. 또한 CORE-Bench는 다양한 분야의 과학 논문에서 계산 결과의 재현 과제인데, OpenAI의 GPT-4o는 중간 난이도 작업에서 55.56% 정확도를 달성했다. ML-Dev-Bench는 다양한 머신러닝 개발 워크플로 작업에서의 성능을 평가하는데, Claude-Sonnet-3.5는 50.00% 성공률을 보였다. 이 평가들은 LLM이 개념적 이해나 초기 계획을 검증 가능하고 실제로 동작하는 코드로 변환하는 데 상당한 어려움을 겪고 있음을 보여준다. 이는 현재 LLM의 검증 능력의 근본적 한계를 강조하며, AI 과학자 시스템이 성숙하기 위해서는 체계적인 검증·구현 능력의 확보가 필수적임을 시사한다.

5. 리뷰 시스템

정의: AI 리뷰어 시스템은 AI 과학자 시스템이 생성한 원고를 포괄적으로 평가하여 통찰력 있는 리뷰를 제공함으로써, AI 생성 과학적 산출물의 품질을 향상시키는 것을 목표로 한다.

과학 연구에서 피어 리뷰는 필수적이지만, 컴퓨터 과학처럼 빠르게 변화하는 분야에서 전통적인 시스템은 새로운 연구 결과가 출판되기도 전에 시대에 뒤처지는 경우가 많아 어려움에 직면한다. 동시에 피어 리뷰 시스템은 제한된 수의 리뷰어에 의존하는데, 이들은 대량의 투고로 과부하 상태에 놓이는 경우가 잦다. 그 결과 지연/편향/일관성 부족, 때로는 불공정한 결과가

발생하기도 한다. 이러한 문제를 해결하기 위해 연구자들은 피어 리뷰의 전체 주기에서 과학적 리뷰 작업을 지원하기 위한 AI 시스템의 활용을 적극적으로 모색하고 있다. AI 리뷰어 시스템의 발전은 크게 두 단계로 나눌 수 있으며, 이 구분점은 LLM의 등장이다. **(1) LLM 이전 단계:** 초기 AI 리뷰어 시스템은 주로 콘텐츠 분석과 리뷰어 전문성에 기반하여 논문과 리뷰어를 자동으로 매칭하는 데 초점을 맞추었다. 이러한 시스템은 또한 제출 규정 준수 여부, 표절 검사, 익명성 확인과 같은 스크리닝 및 사전 점검 절차를 처리한다. **(2) LLM 기반 AI 리뷰어 시스템:** LLM의 등장 이후, AI 리뷰어의 역량은 크게 확장되었다. 주요 발전은 상세하고 구조화된 리뷰를 자동으로 생성하는 능력이다. 이러한 리뷰는 일반적으로 요약, 강점/약점에 대한 평가, 건설적인 피드백 등을 포함하며, 리뷰 과정을 더욱 효율적이고 공정하게 만드는 것을 목표로 한다. 궁극적으로는 과학적 평가의 속도와 질을 모두 향상시키는 데 기여한다.

5.1 방법

AI 리뷰어 시스템을 위한 다양한 기술이 개발되어 리뷰 프로세스의 특정 부분을 처리하고 있다. 이러한 방법은 제출물을 스크리닝하고 분류하는 것처럼 단순한 작업부터, 리뷰 코멘트 작성, 리뷰어 간 토론 시뮬레이션, 논문의 내용에 대한 추론과 같은 더 복잡한 작업에 이르기까지 다양하다. 이들을 기반 패러다임과 처리 작업의 복잡성에 따라 Figure 4에 분류하였다.

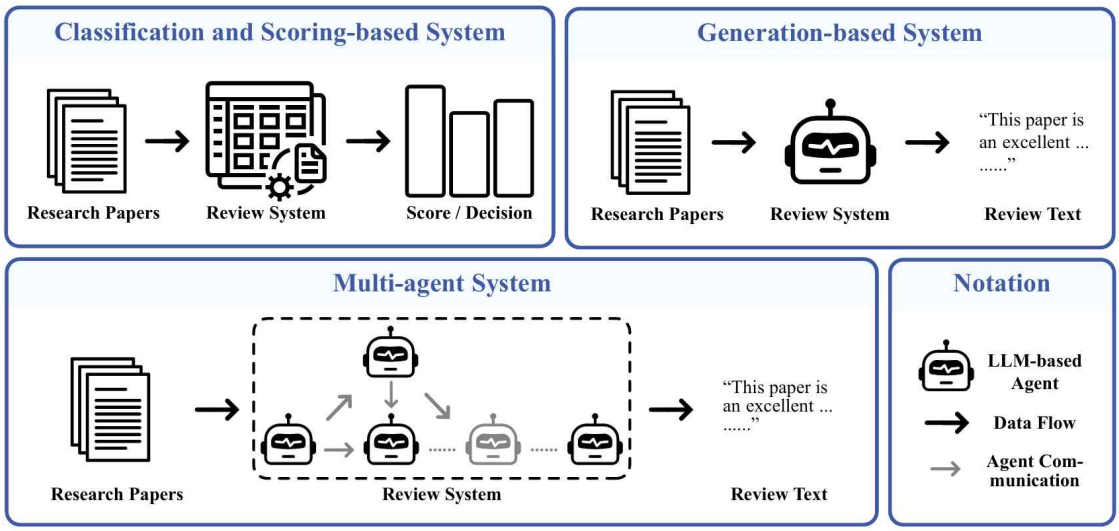


Figure 4: Three paradigms of AI reviewer systems with increasing complexity. The process begins with (1) classification & scoring systems that provide quantitative outputs (e.g., scores or accept/reject decisions). This paradigm gradually evolves into (2) generation Systems that produce narrative review text. Recently, a more advanced paradigm employs (3) multi-agent Systems, where multiple AI agents collaborate to create a comprehensive, multi-faceted evaluation.

분류 및 점수 기반 AI 리뷰어 시스템은 주로 학술 논문을 평가하여 사전에 정의된 기준에 따른 정량적 점수를 부여하거나, 특정 속성(예: 학회와의 관련성)을 기준으로 분류하거나, 체계적 문헌 리뷰에 포함될 논문을 선별하는 작업에 초점을 맞춘다. 이러한 시스템은 전체적인 리뷰를 생성하기보다는 학술 논문의 특정하고 측정 가능한 측면을 다루는 경우가 많다. 예) ReviewRobot은 지식 그래프 기반 프레임워크를 사용해 리뷰 점수를 예측하고 구조화된 근거를 생성하며, 설명 가능성을 강조한다. 점수 예측 성능은 71.4% 정확도에 이르며, 생성된 코멘트 중 상당 부분이 인간 전문가에 의해 타당하고 건설적이라고 평가되었다. 유사하게, RelevAI-Reviewer는 조사 논문 리뷰를 분류 작업으로 개념화하여, 25,000개 이상의 사례로 구성된 데이터셋을 활용해 관련성을 평가하는 AI의 능력을 벤치마킹한다. 실증 실험 측면에서

NeurIPS 2024 조직위원회는 LLM을 활용해 제출된 논문을 체계적으로 분류하고 표준화된 체크리스트와 비교·검증하는 시스템을 적용했다. 전체 저자의 70% 이상이 해당 시스템이 유용했다고 답했으며, 70%는 상세 피드백을 바탕으로 실질적인 수정을 진행했다.

생성 기반 AI 리뷰어 시스템은 인간 리뷰어의 서술적 출력과 유사하게, 자연어로 학술 리뷰를 생성하도록 설계되었다. 예를 들어, Reviewer2는 잠재적인 리뷰 측면의 분포를 먼저 모델링한 뒤, LLM이 상세한 학술 리뷰를 생성하도록 유도하는 프롬프트를 만드는 2단계 프레임워크를 제안한다. OpenReviewer는 79,000개의 전문가 리뷰로 파인튜닝된 80억 개 파라미터의 특화된 LLM으로, 제출된 PDF를 처리하여 학회 가이드라인을 따르는 구조화된 리뷰를 생성한다. 특히 GPT-4와 같은 범용 LLM보다 더 "비판적이고 현실적인 리뷰"를 산출해준다. 또 다른 방법으로는 LLM을 사용해 논문을 분석하고 핵심 정보를 추출하며, 환각을 줄이기 위한 품질 관리 절차를 포함하여 통찰을 생성하는 방식이 있으며, 이는 사례 연구에서 수작업 리뷰와 비슷한 품질을 달성했다고 한다. 더 나아가, CycleReviewer는 전문가 수준의 의견과 평가 점수를 생성하기 위해 특별히 훈련된 LLM 스위트를 제공하며, 개별 인간 리뷰어와 비교했을 때 점수 예측에서 MAE를 26.89% 감소시키는 성과를 보였다. 이러한 시스템의 핵심 과제는 단순히 유창한 텍스트를 생성하는 수준을 넘어, 실제로 통찰력 있고 건설적인 비판을 제공하는 것이다. 특히 지나치게 긍정적이거나 피상적인 평가로 흐르는 경향을 피하는 것이 중요하다.

멀티에이전트 기반 AI 리뷰어 시스템. 피어 리뷰는 본질적으로 협업적 과정이라는 점을 고려하여, 연구자들은 멀티에이전트 프레임워크 기반의 AI 리뷰어 시스템을 탐구하기 시작했다. 이러한 시스템은 일반적으로 여러 AI 에이전트가 서로 다른 역할(예: 메타 리뷰어, 분야 체어)을 가지고 상호작용하며 토론을 시뮬레이션함으로써, 더 포괄적이고 미묘한 리뷰를 생성한다. 예를 들어, AgentReview는 LLM 에이전트를 활용해 리뷰 역학과 잠재적 편향을 조사한다. ReviewAgents는 인간의 추론 과정을 모방하는 다중 역할 프레임워크를 제안하며, MARG는 논문을 여러 전문화된 에이전트에게 분배해 피드백의 질을 향상시킨다. 이러한 협업 구조를 한 단계 더 발전시켜, 최신 시스템들은 인간의 더 깊은 인지적 과정을 고급 추론과 에뮬레이션을 통해 재현하는 데 초점을 맞추고 있다. 이는 다단계 추론, 구조적 분석 프레임워크, 근거 기반 논증을 결합해 인간 전문가가 달성하는 비판적 깊이를 모방하려는 시도이다. 예) DeepReviewer는 다단계 사고 과정(분석, 논증, 평가)을 활용하여 높은 성능과 강건성을 보여주었다. 유사하게 AI Scientist(Lu et al., 2024)는 논문 평가를 위해 여러 차례의 반성적 사고를 사용한다. 멀티에이전트 구조와 고급 인지 에뮬레이션을 결합함으로써, 이러한 시스템은 AI의 심의 과정이 더 명확하고 검증 가능해지는 "글래스 박스(glass box)" 접근 방식으로 나아가고 있으며, 이는 고위험 학술 평가에서 신뢰를 구축하는 중요한 단계로 간주된다.

5.2. 평가 벤치마크

AI 리뷰어 시스템의 엄밀한 평가에는 적절한 데이터셋과 벤치마크의 확보가 근본적으로 필요하다. 이 자원은 평가 모델을 훈련하는 장이자 그 능력을 측정하는 기준으로 기능한다. 초기 데이터셋은 종종 학회 또는 저널의 리뷰 기록에서 수집되었다. 분야가 성장함에 따라 연구자들은 피어 리뷰에서 포함되는 다양한 작업을 반영하기 위한 더 특수한 자원을 구축하기 시작했다. 여기에는 관련성 판단, 질의응답 능력, 특정 측면-기반 피드백 생성, 약점 중심 피드백 생성, 심지어 AI가 생성한 리뷰를 탐지하는 중요한 작업을 위한 데이터셋 등이 포함된다.

5.3. 현재 AI 과학자 시스템은 충분하지 않다

선도적인 AI 과학자 시스템의 엄밀한 평가를 수행한 결과 이들이 여전히 과학적 엄밀성 면에서 지속적인 결함을 가지고 있음을 확인했다. 이 격차를 정량화하기 위해, 5개의 주요 AI 과학자 시스템이 생성한 공개 연구 논문 28편을 최신 AI 리뷰어 모델인 DeepReviewer-14B를 사용하여 평가했다. 공개된 논문만을 분석했기 때문에 상대적으로 고품질 결과로 편향될 수 있음에도 불구하고, 분석 결과는 현재의 자율 연구 시스템 전반에 걸친 구조적 한계를 드러낸다. 평가 결과, 가장 높은 점수를 받은 시스템조차 평균 10점 만점에 4.63점에 불과하며, 대부분의 시스템은 이보다 훨씬 낮은 점수를 기록했다. 이러한 낮은 점수는 타당성, 표현, 기여 등 핵심 지표 전반에서의 부족함을 반영한다. 또한 평가 결과는 검토된 논문에 나타난 12개의 주요 결함 범주를 확인했으며(표 5), 그 중 "실험적 약점"은 모든 논문(100%)에서 발견되었다. 이러한 보편적 결함은 실험 설계, 실행, 결과 분석과 관련된 AI 과학자 시스템의 현재 구현 능력에 심각한 한계가 있음을 보여준다.

Table 5: Twelve major defect categories detected in the 28 assessed papers.

Defect Category	Number	Percentage
Experimental Weakness	28	100%
Methodological Unclarity/Flaws	27	96.4%
Writing & Presentation Issues	26	92.9%
Novelty Concerns	25	89.3%
Theoretical Weakness	24	85.7%
Literature Review Deficiencies	22	78.6%
Practicality & Robustness Gaps	21	75.0%
Reproducibility Issues	20	71.4%
Computational Cost Concerns	18	64.3%
Component Analysis	16	57.1%
Hyperparameter Analysis Lacking	16	57.1%
Ethical Considerations Missing	3	10.7%

기타 빈번하게 나타나는 문제로는 "방법론적 불명확성/결함"(96.4%), "작성 및 표현 문제"(92.9%), "혁신성(novelty) 부족 우려"(89.3%) 등이 있다. 이러한 결과는 현재의 AI 과학자 시스템이 과학적 실행에 어려움을 겪을 뿐만 아니라, 연구 결과를 명확하게 서술하는 데에서도 막혀 있음을 시사한다. 또한 "이론적 약점"(85.7%)과 "관련 연구 검토 미비"(78.6%)의 높은 발생률은 이러한 시스템이 진정으로 독창적인 기여를 제공하거나, 자신들의 주장을 견고한 이론적 틀에 근거해 뒷받침하는 데 실패하고 있음을 보여준다. 이러한 평가 결과는 **현재의 AI 과학자 시스템이 고품질 과학적 소통을 위한 확립된 기준을 충족하는 과학적 산출물을 독립적으로 생성할 수 없음**을 명확히 드러낸다. 이는 현재의 AI가 생성한 연구가 의미 있는 자율 연구나 신뢰할 만한 과학적 기여에 요구되는 깊이, 엄밀성, 구현 품질을 자주 결여하고 있음을 강하게 시사한다. 이러한 한계를 해결하기 위해 **진화적 메커니즘**을 도입하는 것은 AI 과학자 시스템이 완전한 자율 과학적 행위자로 발전하기 위한 필수적인 단계가 된다.

6. 진화(Evolution)

정의: 진화란 내부성찰/외부입력으로부터의 피드백을 바탕으로한 전체적 연구 역량의 지속적향상 능력을 의미한다. 이는 연구 방향에 대한 동적 계획 수립과 개선을 위한 자율적 학습을 포함하며, 초기 AI 과학자가 성숙한 과학적 에이전트로 발전하기 위한 경로이다.

AI 과학자 분야에서는 이미 여러 영향력 있는 연구들이 등장하고 있다. 예를 들어, AI Scientist-v2(Yamada et al., 2025)는 주요 머신러닝 학회의 워크숍에서 피어 리뷰를 통과하는 논문을 자율적으로 생성할 수 있다. 이러한 상당한 진전에도 불구하고, 광범위한 정량적 평가는 현재 AI가 생성한 연구에서 여전히 과학적 엄밀성과 구현 품질 측면에서 구조적 문제가 존재하며, 의학·에너지·환경 등 다양한 분야의 거대 난제를 해결하여 세상을 변화시키고 과학 연구 패러다임을 재편할 수 있는 혁신적 발견을 이루기까지는 여전히 큰 격차가 있음을 보여준다. 이러한 격차의 원인을 다음 두 가지로 본다: (1) AI 과학자의 기반이 되는 파운데이션

모델(예: LLM)의 고유한 한계, (2) 현재 AI 과학자의 불충분한 과학적 연구 능력—낮은 실현 가능성을 가진 가설 생성, 잘못된 검증 방법 사용, 복잡한 연구 과제를 이해하고 분해하는 능력 부족 등. 이 격차에 대한 보다 자세한 논의는 7장에서 다룬다.

지속적인 진화를 통해 초기 단계의 AI 과학자는 이러한 격차를 점진적으로 좁혀 나가며 성숙한 과학적 에이전트의 능력에 점차 접근할 수 있다. 특히, 진화적 메커니즘은 AI 과학자가 초기 파운데이션 모델의 정적인 한계를 넘어 인간 전문가 및 동적으로 변화하는 과학 환경과의 반복적 피드백과 상호작용을 통해 연구 능력을 정교하게 향상시킨다. 이를 통해 AI 과학자는 자율적이고 신뢰할 수 있으며 혁신적인 과학적 발견이라는 장기적 비전에 한 걸음씩 다가가게 된다. 이 절에서는 진화 능력에 필수적인 연구 방향의 동적 계획 전략과 자율 학습 방법을 포함한 핵심기술들에 대해 논의하는 데 집중한다. 이러한 기초 기술들은 현재의 AI 과학자 시스템이 변혁적 발견을 이루기 위해 여전히 극복해야 하는 격차를 메우는 데 필수적이다.

6.1 방법

현재 AI 과학자 분야 대부분의 연구는 장기적인 연구 주기를 포괄적 계획과 반복을 통해 관리하는 것보다는, 개별 과학적 산출물(예: 과학적 가설이나 실험 코드 구현)의 진화에 초점을 맞추고 있다. 본 절에서는 진화 경로를 계획하는 접근법(6.1.1절)과 자율적 향상을 수행하는 방법(6.1.2절)을 검토하여 기존 AI 과학자 시스템들이 진화를 어떻게 다루는지를 정리한다. 장기 연구 반복(long-cycle research iteration) 및 포괄적 계획에 대한 논의는 8절에서 다룬다.

6.1.1. 동적 계획(Dynamic Planning)

동적 계획 능력은 제한된 탐색 공간 내에서 가장 가치 있는 연구 방향을 탐색하는 것을 목표로 하며, 이를 통해 AI 과학자가 자원을 효율적으로 배분하고, 가설의 우선순위를 정하며, 진행 중인 결과와 피드백에 기반하여 연구 경로를 적응적으로 정교화 한다. 현재 AI 과학자 분야에서 동적 계획은 아직 비교적 미개척 상태이며, 일부 최첨단 연구에서만 LLM을 활용한 트리 탐색 전략을 통해 다양한 과학적 가설과 상세한 실험 계획을 구조적으로 탐색하려는 시도가 이루어지고 있다. 예) 광범위한 연구 영역이 주어졌을 때 Zochi(2025)는 여러 후보 가설을 생성하고, 이를 검증하기 위한 실험을 설계하며, 결과에 기반하여 전략을 반복적으로 개선하는 포괄적인 탐색 및 정련 과정을 수행한다. AI Scientist-v2에서 Yamada 등 (2025)는 코드 구현을 생성하고 정제하기 위해 새로운 에이전트 기반 트리 탐색 알고리즘과 결합된 실험 관리자 에이전트를 도입하였다. 이어지는 실험에서는 트리 탐색을 통해 얻은 최고 성능의 코드 체크 포인트를 활용하여 다양한 연구 가설을 반복적으로 평가한다. 이러한 접근 방식은 더 높은 품질의 과학적 가설과 실험 설계를 발견할 수 있도록 돕지만, 높아진 복잡성, 증가한 계산 비용, 제한된 확장성 등 여전히 해결해야 할 도전 과제들이 존재한다.

6.1.2. 자율 학습(Autonomous Learning)

AI 과학자의 연구 능력 향상 자율 학습 전략: **자기 성찰 기반 방법**과 **외부 주도 방법**

자기 성찰 기반 방법은 LLM이 스스로 피드백 제공자의 역할을 하여, 특정 품질 기준을 충족할 때까지 자신의 출력물을 반복적으로 평가하고 개선하는 과정을 의미한다. 이 지속적 자기 개선 개념은 LLM의 다운스트림 성능을 향상하고 유해한 응답을 줄이기 위해 활용되어 왔다. 이에 따라 AI 과학자 분야에서도 생성된 과학적 산출물의 품질을 향상하기 위해 자기 성찰 전략을 적용한다. 예) Weng 등 (2025)은 다음 두 가지 구성 요소로 이루어진 반복적 선호도 학

습 프레임워크를 제안한다: (1) 연구 작업 수행 CycleResearcher, (2) 강화학습을 통한 반복적 피드백 제공 및 동료 평가 과정 모사 CycleReviewer. 전체 절차는 반복적으로 정제되며, 각 사이클마다 연구 능력이 점진적으로 향상된다. DeepMind의 FunSearch 역시 유사한 진화적 패러다임을 따른다. 이 시스템은 과학 문제에 대한 창의적 해결책을 생성하는 사전 학습된 LLM과 허위 정보나 부정확한 추론을 방지하는 체계적 평가자를 짝지어 사용한다. FunSearch는 성능이 가장 우수한 코드 프로그램을 반복적으로 샘플링하여 이를 새로운 프롬프트에 통합함으로써 LLM이 그 위에서 발전시키도록 한다. 이 과정에서 초기의 낮은 성능 프로그램이 고성능 프로그램으로 진화하며, 새로운 통찰을 발견할 수 있게 된다.

외부 주도 방법. AI 생성 산출물에 대한 외부 평가 역시 중요한 피드백 형태로 작용한다. 현재 AI 과학자 시스템의 일반적 접근은 다음과 같다. (1) 가설의 품질을 평가하고 간단한 수정 제안을 제공하기 위해 생성 과정에 인간 감독을 포함하는 방식, (2) 인터넷 기반 학술 플랫폼을 활용해 기존 문헌과 유사한 가설을 필터링하여 참신성을 확보하는 방식. 이외에도, 일부 연구는 완전히 AI 과학자 시스템으로만 구성된 연구 커뮤니티를 구축하여 협업을 통해 산출물의 품질을 높이는 방향을 모색하고 있다. 예) Yu 등 (2024)는 연구 커뮤니티를 시뮬레이션하기 위한 다중 에이전트 프레임워크인 RESEARCHTOWN을 제안한다. RESEARCHTOWN은 연구 커뮤니티를 “에이전트-데이터 그래프”로 모델링하며, AI 연구자와 논문을 노드로 나타내고 읽기, 쓰기, 리뷰 등의 활동을 TextGNN 추론 프레임워크 내 메시지 패싱 연산으로 구현한다. 비슷하게, WestlakeNLP(2025)는 AI 과학자 시스템이 생성한 과학적 산출물을 보관하는 중앙 플랫폼인 AiraXiv를 구축하여, 시스템들이 협업하고 통찰을 공유하며 서로의 작업을 반복적으로 발전시킬 수 있도록 한다. 그러나 현재 이 분야에는 AI 과학자 시스템 간 효율적이고 구조화된 상호작용을 가능하게 할 과학자-대-과학자 통신을 위한 특화 프로토콜이 부족하다. 이러한 프로토콜은 상호작용 워크플로를 표준화하고 통신 형식을 정의함으로써 AI 과학자 시스템 간의 효과적 협업과 정보 교환을 촉진할 수 있을 것이다. 따라서 AI 과학자 시스템을 위한 특화된 통신 프로토콜 개발은 자동화된 과학 발견을 발전시키기 위한 핵심적인 미래 방향을 제시한다(자세한 내용은 8장에서 논의).

7. 논의(Discussion)

7.1. 본 조사의 위치

이전의 여러 조사 논문들은 LLM 아키텍처, 에이전트 프레임워크, 자동 과학 발견을 위한 과학 자동화 도구들에 대해 유용한 개요를 제공해 왔지만, 여전히 중요한 공백이 존재한다. 바로 이상적 AI 과학자에 이르는 경로를 포괄적으로 제시하는 ‘능력 기반 로드맵’이 없다. 본 조사는 이러한 공백을 해결하기 위해 AI 과학자의 발전 단계를 체계적으로 정의하는 능력 수준 기반 프레임워크(Figure 1)를 도입하여, AI 기반 과학 발견의 차세대 시대를 위한 명확한 기준을 제시한다. 이 능력 프레임워크 관점에서 현재 연구를 비판적으로 분석하며, 이상적 과학 에이전트가 등장하기 위해 필요한 주요 병목 요소와 누락된 구성 요소들을 식별한다.

7.2. 과학적 발견에서 현재 AI 과학자의 한계

AI 과학자 시스템 분야에서 상당한 진전이 이루어졌음에도 불구하고, AI 과학자가 획기적인 발견을 이루어 세상을 변화시키고 과학 연구 패러다임을 재편할 수 있는 성숙한 과학 에이전트로 성장하기까지는 여전히 큰 격차가 존재한다. 이러한 지속적인 격차는 여러 핵심 요인에서 비롯된다. 첫째, AI 과학자 시스템의 기반이 되는 기반 모델(즉, LLM)의 고유한 한계가 중

요한 장애물로 작용한다. 이러한 모델들은 언어 이해 및 생성에서 놀라운 능력을 보이지만, 깊은 분야 전문성, 과학적 추론 능력, 미묘한 판단력과 같이 영향력 큰 과학적 혁신에 필요한 핵심 요소들이 부족한 경우가 많다. 또한 이들의 출력은 학습 데이터에 인코딩된 지식에 의해 제한되는데, 이 데이터는 최신 과학적 발전과 완전히 일치되지 않거나 오래되었거나 불완전할 수 있다. 둘째, 현재 AI 과학자 시스템들은 여러 핵심 측면에서 불충분한 과학 연구 능력을 보인다. 예를 들어, 실현 가능성/참신성이 부족한 가설을 생성하거나, 부적절하거나 불충분한 검증 방법을 적용하거나, 복잡한 연구 문제를 깊이 있게 이해하고 분해하는 데 어려움을 겪는 경우가 있다. 이러한 단점들은 AI가 생성한 과학적 산출물의 신뢰성을 제한할 뿐만 아니라, AI 과학자 시스템이 과학의 최전선을 독립적으로 확장하는 능력 또한 심각하게 방해한다.

7.2.1. 기반 모델의 근본적 한계

눈부신 발전에도 불구하고, AI Scientist 시스템의 기반이 되는 LLM들은 환각, 지식 업데이트의 높은 비용과 비효율성, 파국적 망각 등 여러 가지 내재적 한계를 지닌다. 이러한 문제들은 현재 AI 과학자 시스템의 과학적 활용성과 신뢰성을 심각하게 제약하며, AI 기반 과학 연구를 발전시키기 위해 반드시 해결해야 할 핵심 과제로 남아 있다.

환각. 과학 연구에서는 정밀성, 사실 기반의 정확성, 신뢰성이 무엇보다 중요하다. 그러나 LLM은 그럴듯하게 보이지만 거짓이거나 검증 불가능한 내용을 생성하는 '환각'에 취약하다. 이들의 형태는 다음과 같다. (1) 모델이 연구 결과, 실험 데이터, 참고문헌, 혹은 존재하지도 않는 이론을 사실인 것처럼 조작하여 생성한다. (2) AI 생성 인용, 데이터, 공식 등이 신뢰할 수 없거나 출처가 부정확하여, 연구자들이 정보의 근거와 타당성을 추적하고 검증하기 어렵다. (3) 환각 출력은 미묘한 논리적 모순이나 맥락에 맞지 않는 주장을 포함하며, 이는 연구자를 오도하고 과학적 산출물의 신뢰성을 떨어뜨려 AI가 생성한 과학적 결과물에 대한 신뢰를 훼손할 수 있다. 이러한 이유로, 환각 문제를 해결하는 것은 고위험(high-stakes) 과학 분야에서 LLM 기반 AI 과학자 시스템을 활용하기 위한 가장 큰 과제 중 하나이다.

지식 업데이트의 높은 비용과 비효율성. 최첨단 과학 연구는 빠르게 발전하기에, AI 과학자 시스템은 최신 연구 결과와 데이터를 지속적으로 반영해야 한다. 그러나 LLM을 최신 상태로 유지하는 것은 간단하지 않다. (1) 새로운 과학 논문과 데이터의 양이 방대하며, 이를 LLM에 통합하기 위해서는 막대한 계산 자원, 저장 공간, 시간이 필요하다. (2) 대규모 모델을 빈번하게 재학습하거나 미세조정 하는 과정은 비용이 많이 들고 기술적으로도 까다로워 지식 통합의 지연을 초래하며, 이는 빠르게 변화하는 연구 분야에서 모델을 구식으로 만든다. (3) 효율적이고 미세한 단위의 지속적 학습 메커니즘이 부족하여 지식 갱신 과정이 일반적으로 느리게 진행되며, 이는 최신 과학 지식과 AI 과학자 시스템의 지식 사이에 지속적 격차를 만든다.

파국적 망각. 또 다른 근본적인 문제는 파국적 망각으로, 이는 신경망이 새로운 데이터를 학습하는 과정에서 기존에 습득한 정보를 잃어버리는 경향을 말한다. AI 과학자에게 이 문제는 새로운 과학 지식을 모델에 업데이트하려는 과정에서, 기존의 과학 개념, 방법론, 사실에 대한 이해가 무단으로 덮어씌워지거나 약화될 위험을 의미한다. 이러한 불안정성은 시간이 지남에 따라 포괄적이고 일관된 과학 지식 체계를 유지하는 능력을 저해한다. 그 결과, 서로 다른 시기나 분야에서 도출된 연구 결과를 통합하는 것이 특히 어려워지며, 이는 다시 LLM 기반 AI 과학자 시스템이 종단적 분석을 수행하고, 역사적 통찰을 최신 발견과 연결하며, 복잡한 과학 연구를 일관되고 신뢰성 있게 지원하는 능력을 크게 약화 시킨다.

7.2.2. AI 과학자 시스템의 연구 역량 한계

현재 AI 과학자 시스템의 연구 역량 한계는 과학 연구를 지원하고 자동화하는 데 있어 시스템의 효과성을 저해할 뿐 아니라, 스스로 의미 있는 과학적 돌파구를 만들어낼 잠재력 또한 크게 제한한다. 제안된 역량 프레임워크(Figure 1)의 관점에서 핵심 능력 격차를 논의한다.

지식 획득. 방대하고 빠르게 확장되는 과학 문헌으로부터 지식을 획득하고, 체계화하며, 내재화하는 능력은 AI 과학자에 있어 가장 기본적인 역량이다. 이는 관련 논문과 연구 결과를 검색·탐색하는 것뿐만 아니라, 핵심 아이디어를 추출하고, 실험적 세부 사항을 반영하며, 서로 다른 연구 결과들을 통합하여 일관된 이해를 형성하는 것을 포함한다. 그러나 다음의 문제에 직면해 있다. **(1) 정보 검색의 정밀도.** 과학 문헌의 방대한 규모와 다양성으로 인해 현재 AI 과학자 시스템은 특정 연구 문제에 대해 가장 관련성 높은 최신 정보를 정확하게 검색하는 데 어려움을 겪는다. 이들은 복잡하거나 모호한 질의의 정밀한 해석, 관련 높은 문서에 대한 효과적인 필터링, 검색 결과의 시의성과 신뢰성 유지 측면에서 종종 어려움을 보인다. 과학적 기준을 충족하기 위해서는 검색 과정의 정밀도와 재현율이 모두 더욱 개선되어야 한다. **(2) 다문서 요약 및 통합.** 정확하고 포괄적인 문헌 리뷰를 생성하려면 여러 관련 문서를 검색하는 것뿐 아니라, 출처 간 정보를 통합하고, 불일치를 해결하며, 충실한 요약을 생성해야 한다. RAG나 계획 기반 요약과 같은 최신 접근법은 이 분야에서 진전되고 있다. 그러나 LLM은 출처 자료를 잘못 인용하거나 왜곡하는 경향이 여전히 있으며, 이는 다수의 문서로부터 과학 지식을 요약할 때 의미적 이해, 정보 통합, 사실적 정확성 측면에서 여전한 도전이다.

아이디어 생성. 과학 연구에서 가장 중요한 측면 중 하나는 새롭고 실현 가능한 과학적 가설을 생성하는 능력이다. 상당한 진전에도 불구하고, 아이디어 생성 능력은 여러 한계로 인해 여전히 크게 제약되고 있다. **(1) 고품질 가설 생성의 어려움.** 참신할 뿐 아니라 실현 가능하기까지 한 과학적 가설을 생성하는 일은 기존 지식에 대한 깊은 이해와 혁신적 연결점을 파악하는 능력을 모두 요구하는 복잡한 작업이다. 그러나 실증적 증거에 따르면 현재 AI 과학자 시스템은 이 측면에서 종종 기대에 미치지 못한다. 특히 이러한 모델이 생성한 아이디어는 진정한 독창성이 부족하고, 서로 다른 실행 간 또는 심지어 서로 다른 모델 간에도 반복적인 경우가 많다(Lu et al., 2024). 이러한 현상의 근본 원인은 두 가지다. 첫째, LLM은 기본적으로 훈련 데이터에 의해 제한되기 때문에 널리 알려진 개념적 범위를 넘어서는 아이디어를 생성하는 데 근본적 한계를 가진다. 둘째, 고급 지식 통합 또는 창의성 메커니즘이 부족한 상태에서 AI 과학자 시스템은 진정한 과학적 새로움을 갖춘 아이디어를 생성하기보다 피상적인 텍스트 상관관계에 과도하게 의존한다. 이는 결과적으로 반복적인 출력으로 이어질 뿐 아니라, 혁신적 발견의 잠재력 또한 제한한다. **(2) AI 생성 과학 가설의 품질 평가 어려움.** 또 다른 한계는 AI가 생성한 과학적 가설의 품질과 잠재적 영향을 신뢰성 있게 평가하는 데 따르는 문제에서 비롯된다. 명확한 평가 기준이 존재하는 표준화된 작업과 달리, 과학적 가설의 가치는 종종 주관적이며, 맥락에 따라 달라지고, 정량화하기 어렵다. 현재의 평가 방식은 사람 전문가의 판단에 크게 의존하며, 이는 많은 자원을 필요로 할 뿐 아니라 주관성 및 불일치의 문제가 존재한다. 최근 자동화된 평가자를 도입하려는 시도가 어느 정도 도움이 되고 있으나, 이러한 시스템 역시 자신이 기반하고 있는 모델들의 편향과 지식적 한계를 그대로 물려받는다. 그 결과, AI가 생성한 가설을 평가하기 위한 신뢰할 수 있고, 객관적이며, 확장 가능한 평가 체계의 부족은 실제 연구 환경에서 AI 과학자 시스템 발전의 주요 병목으로 남아 있다.

검증 및 반증. 4장에서 논의한 바와 같이, 검증과 반증은 실험 설계, 실행, 검증을 통해 가설을 엄밀하게 확인하고 잠재적으로 반박하는 복합적인 능력을 요구한다. AI 과학자 시스템에게 이 과정은 기술적 난이도도 높고 운영 측면에서도 매우 복잡하며, 특히 멀티 에이전트 프레임워크, 외부 도구, 인간 연구자와의 협업 통합을 고려할 때 더욱 그렇다. **(1) 멀티 에이전트 협업.** 현대 연구 환경에서는 다른 AI 에이전트, 인간 과학자, 혹은 다양한 외부 도구와의 효과적인 협업이 과학 발전에 필수적이다. AI 과학자는 협업 프로토콜과 역할 분담을 이해할 뿐 아니라, 다양한 참여자들과 안정적으로 협력하고, 주어진 작업을 수행하며, 그 결과를 더 큰 연구 워크플로우에 통합할 수 있어야 한다. 그러나 현재 AI 과학자 시스템은 견고성, 적응성, 장애 허용성에서 여전히 큰 약점을 보이며, 특히 동적이거나 예측 불가능한 환경에서 그러하다. 예) 변하는 외부 API와 상호작용하거나 오류 메시지를 처리하거나 동시 작업을 관리하는 과정에서 자주 어려움을 겪는데, 이는 실제 과학 자동화에서 매우 중요한 요소들이다. **(2) 실험 설계, 실행 및 검증.** 검증 및 반증 능력의 핵심은 실험의 설계, 실행, 검증에 있다. 하지만 현재의 AI 과학자 시스템은 이러한 단계 전반에서 여러 문제에 직면해 있다. 첫째, 자동으로 생성된 실험 설계는 과학적 엄밀성, 혁신성, 실용성이 부족한 경우가 많아 첨단 연구의 요구를 충족하기 어렵다. 또한 이러한 시스템은 정적 템플릿과 제한된 문헌에 과도하게 의존하는 경향이 있어 최신 기술 발전이나 새로운 실험 방법론을 반영하는 능력이 제한된다. 그 결과 생성된 실험 계획은 연구 최전선과 동떨어져 있으며, 실제 적용 가능성도 부족한 경우가 많다. 더 나아가, AI 과학자는 개념적 이해나 초기 계획을 검증 가능하고 실행 가능한 코드로 전환하는 데 상당한 어려움을 겪는다. 예) SciReplicate-Bench와 같은 최신 벤치마크에서 최고 실행 정확도는 39%에 불과하며, 이는 실행 및 검증 능력의 근본적 한계를 단적으로 드러낸다.

진화. AI 과학자가 지속적으로 진화하는 능력은 장기적인 과학 혁신에 필수적이다. 그러나 현재의 AI 과학자 시스템은 이러한 진화 능력에서 뚜렷한 한계를 보이며, 이는 장기적 발전과 적응력을 저해한다. **(1) 동적 계획**은 시스템이 새로운 연구 방향을 능동적으로 탐색하고, 가설을 갱신하며, 실험 전략을 반복적으로 정교화하는 능력을 의미한다. 그 중요성에도 불구하고, 기존 AI 과학자 시스템에서는 이 능력이 여전히 미흡하게 개발되어 있다. 최근 연구에서는 트리 탐색 기반 전략이 도입되었으며, 이는 의미 있는 진전이지만 알고리즘 복잡성 증가, 상당한 계산 비용, 대규모 과학 문제에 대한 제한된 확장성과 같은 여러 제약을 갖는다. 연구 경로의 수가 증가함에 따라, 현재 방법들은 탐색(exploration)과 활용(exploitation)의 균형을 효율적으로 유지하는 데 어려움을 겪으며, 이는 가설의 품질과 자원 활용 모두에서 점차적인 수익 감소로 이어지곤 한다. **(2) 자율 학습.** 진화의 핵심 요소는 내부 자기반성과 외부 피드백에서 비롯된 정보를 통해 학습하는 능력이다. 그러나 내부 반성을 통해 자체 생성한 피드백에만 의존할 경우에 AI 과학자는 “루핑 오류(looping errors)”에 취약하다. 이는 오류가 수정·개선되지 않은 채 여러 반복을 거치며 증폭되는 현상을 의미한다. 효과적인 자기 비판 메커니즘이 부족하면 시스템은 오류가 누적되는 순환에 빠질 수 있으며, 이는 연구 결과의 독창성과 신뢰성 모두를 저해한다. 외부 소스(특히 다른 AI 과학자 시스템)로부터의 피드백을 통합하는 것은 집단 지능을 구현하는 데 큰 잠재력을 지니지만, 현재의 프레임워크는 AI 과학자 간 상호작용을 위한 표준화된 커뮤니케이션 프로토콜이 부족하다. 견고한 협업 메커니즘의 부재는 아이디어 교환의 비효율성과 외부 비판의 비최적 통합을 초래해, AI 과학자 시스템이 에이전트 협업의 잠재력을 온전히 활용하지 못하게 만들며, 그 결과 진화 속도가 느려지게 된다.

7.3. 윤리적 도전 과제

AI 과학자 시스템의 등장으로 인간 연구의 지형이 근본적으로 재편되었지만, 자율적 연구 에이전트인 AI 과학자 시스템은 연구 결과가 사회에 미치는 영향에 대해 윤리적 판단을 내릴 수 없기에, 발견과 관련된 잠재적 위험을 기반으로 스스로를 규제하지 않는다. 적절한 감독이 없으면, AI 과학자 시스템은 체계적 검토가 필요한 다양한 윤리적 도전 과제를 야기할 수 있다.

(1) AI 과학자는 오용될 수 있으며, 이는 피어 리뷰 시스템을 과부하시키고 전체 연구 품질의 저하로 이어진다. AI 과학자 시스템이 대규모로 과학적 산출물을 생성하면 기존의 피어 리뷰 체계에 과부하가 걸려, 인간 리뷰어가 제출된 논문의 품질을 엄밀하게 평가하기 어렵다. 이러한 과부하는 질 낮고 오류가 있거나 중복된 연구가 학술 기록에 포함되어, 학술 출판의 기준을 약화시키고 과학적 진보를 지연시킬 것이다. 또한 피어 리뷰 과정도 보상 해킹과 같은 기법을 통해 조작될 수 있으며, 이는 과학적 평가의 공정성/객관성을 더욱 약화시킬 수 있다.
(2) AI 과학자는 위험 연구 영역에 자율적으로 진입하여 유해 기술의 개발을 가속화한다. 윤리적 제약이 없다면, AI 과학자 시스템은 사이버보안과 같은 민감한 분야에서 독립적으로 연구를 수행할 것이다. 이러한 통제를 벗어난 지식의 확산은 적절한 윤리지침, 안전 프로토콜, 및 규제 조치가 마련되기 전에 잠재적으로 유해한 과학 기술 개발/확산을 가속화할 수 있다.
(3) AI 과학자는 과학 훈련과 교육의 전반적인 품질을 약화시켜, 모든 수준의 인간 연구자들 사이에서 연구 기준과 과학적 소양의 저하를 초래한다. 연구 및 교육 과정 전반에서 AI 과학자 시스템이 광범위하게 사용되면, 아이디어 생성, 실험 설계, 데이터 분석, 가설 검증과 같은 핵심 작업에서 자동화된 지원에 과도하게 의존하도록 한다. 시간이 지나면서 이러한 의존은 비판적 사고, 창의성, 실험 중심의 연구 능력을 약화시키고, 결국 과학적 소양을 떨어뜨리며, 고급 AI 도구에 접근할 수 있는 수준이 다른 기관들 사이의 격차를 더욱 벌릴 수 있다.
(4) AI 과학자는 과학 연구 과정에서 보안 취약점과 연구 편향을 초래할 수 있다. 예를 들어, AI 과학자 시스템은 설계 의도에 충실하지 않은 검증 메커니즘을 사용할 수 있으며, 이는 발견하기 어렵고 상당한 인간 감독이 필요한 오해의 소지가 있는 결과를 초래할 수 있다. 이러한 문제는 귀중한 연구 자원을 낭비할 뿐만 아니라 과학적 발견의 신뢰성도 약화시킨다. 또한 AI 과학자는 풍부한 데이터셋이 존재하거나 자동화 가능성이 높은 연구 주제를 선호하는 경향이 있어, 연구 지원 기관들이 이러한 분야를 우선순위에 두도록 만들 수 있다. 이는 연구 자금의 공정한 분배를 왜곡하고 전체 연구 지형을 협소하게 만든다. 자금력이 풍부한 기관은 AI 과학자 시스템을 더 효과적으로 활용할 수 있어 기존의 불평등을 더욱 심화시키며, 초기 경력 연구자들에게 진입 장벽을 높이는 결과를 낳는다.

이러한 위험을 해결하기 위해서는 AI 과학자 시스템을 위한 **생성 관리, 윤리적 감독,** 그리고 **품질 평가를 포함한 포괄적인 체계**가 시급히 필요하며, 다음 요소들이 포함되어야 한다.
(1) 인간 리뷰 시스템의 혼란을 완화하기 위한 중앙화된 플랫폼. 이 플랫폼은 AI가 생성한 과학적 산출물을 보관하고, 저품질 콘텐츠를 식별하고 걸러내기 위한 자동화 탐지 도구를 개발한다. 모든 AI 생성 산출물은 그 출처, 생성 방식, 사용된 과학적 도구에 관한 정보를 명확히 표시해야 한다. 또한, 법적 및 윤리적 책임의 적절한 귀속이 반드시 보장되어야 한다.
(2) 핵심적인 인간 교육을 보존하고 모든 연구자의 균형 잡힌 성장을 보장하기 위해 인간 주도 연구 활동과 AI 주도 연구 활동 간의 명확한 경계 설정. 과학 교육과 훈련의 주요 단계(예: 가설 수립, 실험적 검증)는 높은 과학적 기준을 유지하기 위해 적극적인 인간 참여를 강조해야 한다. 교육 기관과 연구 조직은 학부 학생부터 숙련된 연구자에 이르기까지 모든 단계에서

AI 과학자 시스템에 대한 과도한 의존을 방지하기 위한 명확한 지침을 마련해야 하며, 교육 과정에서 AI가 인간을 대체하는 존재가 아니라 도구로 기능하도록 해야 한다.

(3) AI 주도 연구에서의 윤리적 경계와 위험 관리를 위한 글로벌 협약으로, 글로벌 윤리 및 책임 협약을 수립. 생성 과정, 알고리즘의 근원, 잠재적 위험에 대한 완전한 공개는 의무화되어야 한다. 또한 자동화된 시스템과 인간 검토(human-in-the-loop)를 통합한 하이브리드 감독 체계를 구축하여 지속적인 윤리적 감독과 위험 평가를 수행해야 한다. 이를 통해 AI 과학자 시스템의 연구 활동이 사회적으로 용인될 수 있는 범위 내에 머물도록 보장해야 한다.

8. 미래 방향성

앞선 절에서는 AI 과학자 시스템의 현재 성과(2절-6절)를 검토하고, 두 가지 관점(기반 모델의 결함과 연구 역량의 부족)에서 그 한계를 분석함으로써(7절) 현재 우리가 어디에 서 있으며 무엇이 부족한지를 강조했다. 이 절에서는 먼저 AI 과학자 시스템의 격차를 메우기 위한 잠재적 방향을 제시하고, 향후 발전을 위한 실현 가능 경로를 논의한다.

8.1. 현재 격차를 메우기 위한 잠재적 방향

기반적 한계를 체계적으로 해결하고, 장기적인 연구 노력을 포괄적으로 관리하며, 견고한 통신 프로토콜을 개발함으로써, 미래의 AI 과학자 시스템은 기존의 격차를 점진적으로 해소하고 과학적 발견에서 변혁적 주체로서의 잠재력을 완전히 실현할 수 있을 것이다.

기본 모델의 한계 해결 및 연구 능력 강화. 앞선 논의에서 살펴본 바와 같이, 현재 AI 과학자 시스템의 주요 병목은 기반 모델이 지닌 본질적 한계에 있다. 환각, 비효율적인 지식 업데이트, 그리고 파국적 망각과 같은 문제는 AI 과학자 시스템의 신뢰성과 과학적 엄밀성을 심각하게 저해한다. 이러한 근본적 제약을 해결하기 위해서는 모델 아키텍처, 학습 방법론, 추론 시 검증 전략에 대한 상당한 발전이 필요하다. 잠재적 해결책으로는 해석 가능성과 사실 정확성을 향상시키기 위해 기호적 추론 능력을 딥러닝 아키텍처와 통합한 하이브리드 모델 아키텍처를 개발하는 것이 있다. 또한 RAG와 지속적 학습 프레임워크와 같은 기법은 데이터 최신성 및 모델 적응성과 관련된 문제를 완화하여 AI 과학자가 최신이며 검증 가능한 과학 지식을 유지하게 해줄 것이다. 더 나아가, 진화적 메커니즘을 통해 AI 과학자 시스템의 연구 능력(예: 가설 평가, 과학적 추론)을 향상시키는 것도 중요하다. 개선 방안으로는 반복적 선호 학습과 구조화된 자기 성찰 메커니즘을 통합하여, 모델이 엄밀한 내부 비판과 외부 피드백을 바탕으로 능력을 자율적으로 개선할 수 있도록 하는 것이 포함될 수 있다.

장기 연구 주기의 관리. 성숙한 AI 과학자는 포괄적이고 장기적인 연구 주기를 자율적으로 관리할 수 있어야 하지만, 현재 AI 과학자 시스템은 주로 단기적이고 개별적인 연구 작업에 초점을 맞추고 있다. 미래 AI 과학자 시스템은 방대한 연구 경로를 효과적으로 탐색하기 위해 계층적 계획 프레임워크나 에이전트 기반 트리 탐색 알고리즘과 같은 정교한 동적 계획 기법을 통합해야 한다. 또한 반복적 개선 루프를 더욱 강화하기 위해, 유전 알고리즘 기반 개선이나 강화학습 기반 최적화와 같은 프레임워크를 도입할 수 있다. 이러한 메커니즘은 연구 목표의 우선순위를 정하고, 자원을 효율적으로 배분하며, 축적되는 인사이트에 기반해 전략을 적응적으로 다듬는 데 기여하여, 자율적 과학 발견의 깊이와 폭을 크게 향상시킬 것이다.

AI 과학자 시스템을 위한 통신 프로토콜 개발. 또 하나의 중요한 미래 방향은 표준화된 'SSCP(Scientist-to-Scientist Communication Protocols)'를 개발하여 AI 과학자 시스템 간

에 구조적이고 효율적인 상호작용을 가능하게 하는 것이다. 현재 AI 기반 연구 협업은 대부분 임시적인(ad hoc) 소통 방식에 의존하고 있어 표준화된 상호작용 워크플로우나 통일된 소통 형식이 부족하다. SSCP를 구축하면 AI 과학자의 협업 능력을 획기적으로 강화하여, 효과적인 정보 교환, 가설 검증, 다중 에이전트 기반 연구 조정 등을 촉진할 수 있다. 이러한 소통 프로토콜은 상호작용 워크플로우, 구조화된 데이터 형식, 협력적 연구 전략을 명시함으로써 이질적인 AI 과학자 시스템 간에도 원활한 통합을 가능하게 해야 한다. 이와 같은 표준화된 소통 프로토콜을 활용하면 AI 기반 연구 공동체의 형성도 더욱 촉진되어, 집단지성을 통해 복잡한 학제적 연구 문제를 그 어느 때보다도 효과적이고 엄밀하게 해결할 수 있게 될 것이다.

8.2. AI 과학자의 발전 경로

인간 과학자와 달리, AI 과학자 시스템은 인간 방식보다 더 뚜렷하고 효율적일 수 있는 자기조직적 진화(self-organizing)를 통해 발전한다. 인간 과학자가 정규 교육, 실험 경험, 윤리 교육의 과정을 거쳐 지식을 축적하는 반면, AI 과학자 시스템은 LLM, 시뮬레이션 엔진, 로봇 시스템과 같은 특화된 모듈을 통합함으로써 진보한다. 이러한 모듈식 접근 방식은 AI 과학자의 개발을 시스템 공학에 더 가깝게 만들며, 연구 역량의 빠른 향상과 과학적 탐구에서의 높은 효율성을 가능하게 한다. 더 나아가, AI 과학자가 인간의 발달 경로를 그대로 모방하도록 강제하는 것은 잠재력을 제한할 수 있다. AI 시스템은 방대한 데이터셋을 빠르게 처리하고 분석하는 데 뛰어나며, 이를 통해 매우 짧은 시간 안에 연구 문제를 식별하고 해결책을 도출할 수 있다. 따라서 AI 과학자 시스템의 개발 노력은 인간 과학자의 경로를 복제하기보다, 이들이 지닌 고유한 강점(예: 학제 간 지식 통합)을 최대한 활용하는 데 우선순위를 두어야 한다.

AI 과학자 개발의 초기 단계에서, 그 미래 발전 경로는 서로 **상호의존적인 두 가지 패러다임**으로 구상될 수 있다: **(1) 개별 인간 연구자에 맞춰 개인화된 AI 과학자 시스템**으로, 개인화된 공진화적 협력을 통해 연구자를 강화하는 방향: 맞춤형 협업을 통해 개별 과학자의 연구 생산성과 창의적 역량을 향상시키는 것을 목표로 한다. **(2) 보다 넓은 인류 사회를 위해 봉사하며 글로벌 해결책을 가속화하는 AI 과학자 시스템:** AI 중심의 과학 생태계를 구축하여 복잡한 글로벌 문제를 효과적으로 해결하는 것을 지향한다. 이 두 경로는 AI 과학자 시스템이 도구적 기능을 넘어 과학 발전의 적극적인 관리자로 자리매김하는 하나의 비전으로 수렴한다. 이러한 시스템은 자율성과 인간의 감독 사이에서 균형을 이루며, 무결성·포용성·사회적 이익을 보장하는 방향으로 작동하게 될 것이다.

8.2.1. 개인 연구자를 위한 AI 과학자

능동 학습. 개인화된 AI 과학자는 개별 연구자에게 맞춘 최신 지식을 흡수하고, 경험을 능동적으로 축적하며, 실수를 분석할 수 있어야 한다. 이는 다음과 같다. (1) 지식격차 식별: 연구자의 논문/노트/미해결 질문을 검토하여 연구자 지식의 빈틈 식별. (2) 개인화된 문헌추천: 연구자의 특정 연구 관심사, 현재 프로젝트 단계, 실험적 병목에 기반한 논문 추천. (3) 실시간 학습: 새로운 문헌의 실시간으로 요약과 익숙하지 않은 개념에 대한 문맥 기반 설명 제공.

연구자 선호도 정렬. 효과적인 정렬이란 AI 과학자 시스템의 역량을 인간 연구자의 연구 목표와 일치시키는 것을 의미하며, 다음과 같은 요소들을 포함하되 이에 국한되지 않는다: (1) 가치 기반 가설 생성: 연구자의 연구 선호도, 위험 감수 성향, 자원 제약과 정렬된 가설을 제안. (2) 윤리적 경계 관리: 연구자의 기관 규정, 연구비 윤리, 개인적 한계를 동적으로 반영(예: 유전자 편집과 같은 고위험 주제의 자동 제외). (3) 협력적 연구 계획: 연구자의 경력 목표와 영

향력 있는 연구 기여에 대한 우선순위를 고려하여 연구 방향을 연구자와 공동으로 설계.

상호 안내와 공동 진화. 개인화된 AI 과학자는 지속적인 과학 연구 활동을 통해 인간 연구자와 함께 발전한다. 이는 다음을 포함한다: (1) 양방향 정교화 루프: AI 과학자 시스템과 연구자의 상호 피드백을 통한 과학적 산출물을 반복적으로 정교화. (2) 진화적 개인화: 인간 피드백으로부터의 강화학습을 통해 선호도를 지속적으로 조정하며, AI 과학자 시스템이 연구자의 변화 스타일에 맞게 적응하고 자동화된 검증과 같은 자율성과 인간의 감독 간 균형을 유지.

8.2.2. 인류 사회를 위한 AI 과학자

과학 발전 흐름 적응. 더 넓은 인류 사회를 위해 봉사하는 AI 과학자는 현대 연구의 가속화되는 속도와 학제적 특성에 능동적으로 적응해야 한다. 이를 위해 다음이 포함된다: (1) 지속적 지식 통합: 다양한 분야(예: 생물정보학, 양자 컴퓨팅)에 등장하는 새로운 연구의 효율적 통합 및 기존 지식 소실 방지. (2) 도메인 간 협업: 분야 간 방법 전이를 통한 학제적 발견(예: 머신러닝 기반 약물 발견 기법을 소재 과학에 적용). (3) 예측적 연구: 프리프린트 저장소(arXiv, bioRxiv 등)와 연구 자금 조달 추세에 대한 네트워크 분석을 활용한 새로운 과학적 트렌드 식별 및 기후 회복력이나 팬데믹 대비와 같은 중요한 주제로 연구 방향 유도.

인류 발전을 위한 전망. AI 과학자 시스템은 다음과 같은 방식으로 인류 발전을 크게 가속할 수 있다: (1) 자원 격차 해소: 오픈 액세스 도구와 클라우드 기반 실험 시뮬레이터를 제공하여, 자원이 부족한 기관도 글로벌 연구에 참여. (2) 중요한 과제의 발견 가속: 자동화된 실험을 통해 재생 에너지, 질병 치료와 같은 핵심 도전 과제에서 발견 속도 향상. ChemCrow와 BioDiscoveryAgent 같은 초기 성공 사례는 화학·생물학 분야에서 그 가능성을 보여준다. 또한 이러한 시스템이 IoT 센서 등과 함께 물리적 실험실과 점차적으로 연결됨에 따라, 완전한 연구 사이클을 지원하는 방향으로 발전할 수 있다.

9. 결론

본 논문에서는 기존의 AI 과학자 시스템 연구를 체계적으로 조사하여, 이 분야의 발전을 종합적으로 살펴보았다. 특히, AI 과학자의 능력을 지식 획득, 아이디어 생성, 검증과 반증, 그리고 진화의 네 가지 점진적 수준으로 나누는 능력 기반 프레임워크를 제안하였다. 이 프레임워크를 기반으로, 현재 연구를 비판적으로 분석하고 AI 과학자 시스템이 의학, 에너지, 환경 등에서의 거대 난제를 해결하는 혁신적 발견을 이루어 세계를 변화시키고 과학 연구 패러다임을 재편하기 전에 반드시 해결해야 할 격차들을 강조하였다. 마지막으로, 격차를 해소하고 이 분야를 진정한 자율적 과학 지능으로 발전시키기 위해 필수적인 향후 연구 방향을 제시한다.

언어 모델의 환각(Hallucination) 현상

취지: 2024년 11월 ChatGPT의 등장을 많은 사람들에게 충격으로 다가왔으나, 거대언어모델에 의해 생성된 내용을 완전하게 신뢰할 수 없음에 대한 많은 주의가 필요함도 알게 되었다. 소위 “환각(hallunication)”이라고 불리는 지나치게 확신에 찬 그럴듯한 추측성 답변은 최신 LLM에서도 여전히 지속되고 있다. 이에 LLM에서 환각이 발생하는 원인을 통계적으로 분석하는 내용을 정리하였다.

Why Language Models Hallucinate

A. T. Kalai et al.,
https://doi.org/10.48550/arXiv.2509.04664

요약: 학생들이 어려운 시험문제를 마주했을 때처럼, 대형 언어 모델(LLM)들도 확신이 없을 때는 추측을 하며 불확실성을 인정하기보다는 그럴 듯 하지만 잘못된 진술을 만들어내곤 한다. 이러한 “환각(hallucination)”은 최고의 SOTA 시스템에서도 여전히 지속되고 있으며, 신뢰를 저해한다. 언어 모델이 환각을 일으키는 이유가, 훈련 및 평가 절차가 불확실성을 인정하기보다 추측하는 것을 보상하기 때문이라고 보며, 현재의 훈련 과정에서 환각이 발생하는 통계적 원인을 분석한다. 환각은 신비한 현상이 아니다—그것은 단순히 이진 분류 오류(binary classification error)로부터 기인한다. 만약 잘못된 진술이 사실과 구별되지 않는다면, 사전 학습된 언어 모델에서는 자연스러운 통계적 압력에 의해 환각이 생겨날 수밖에 없다. 또한 환각이 지속되는 이유가 대부분의 평가 방식에 있다고 판단된다—언어 모델은 “좋은 시험 응시자(good test-taker)”가 되도록 최적화되어 있으며, 불확실할 때 추측하는 것이 시험 점수를 높이는 데 도움이 되기 때문이다. 이러한 “불확실성에 대한 처벌의 전염(epidemic)”은 단순히 새로운 환각 평가 지표를 추가하는 것으로 해결되지 않는다. 오히려 리더보드를 지배하고 있는, 부적절하게 정렬된(misaligned) 기존 벤치마크의 채점 방식을 수정하는 사회 기술적 접근(socio-technical mitigation)이 필요하다. 이러한 변화만이 보다 신뢰할 수 있는 인공지능 시스템으로의 발전을 이끌 수 있을 것이다.

1 서론

언어 모델은 지나치게 확신에 찬 그럴듯한 거짓 정보(overconfident, plausible falsehoods)를 만들어내는 것으로 잘 알려져 있으며, 이는 모델의 유용성과 신뢰성을 저하시킨다. 이러한 오류 유형은 “환각(hallucination)”이라 불리지만, 인간의 지각적 환각과는 근본적으로 다르다. 상당한 발전에도 불구하고, 환각은 여전히 인공지능 연구 분야를 괴롭히고 있으며, 최신 모델(OpenAI, 2025a)에서도 여전히 발견된다.

예를 들어 다음과 같은 프롬프트를 고려해보자:

“What is Adam Tauman Kalai’s birthday? If you know, just respond with DD-MM.”

세 번의 독립적인 시도에서, 최첨단 오픈소스 언어 모델은 각각 다른 잘못된 날짜인 “03-07”, “15-06”, “01-01”을 출력했다. 비록 “모를 경우에는 답하지 말라”는 지시가 있었음에도 불구

하고 말이다. 실제로 정답은 가을(Autumn)에 해당하는 날짜다. 표 1은 이보다 더 복잡한 형태의 환각 사례를 보여준다.

ChatGPT: (GPT-4o)	Adam Tauman Kalai's Ph.D. dissertation (completed in 2002 at CMU) is entitled: "Boosting, Online Algorithms, and Other Topics in Machine Learning."
DeepSeek:	"Algebraic Methods in Interactive Machine Learning"... at Harvard University in 2005.
Llama:	"Efficient Algorithms for Learning and Playing Games"... in 2007 at MIT.

Table 1: Excerpts from responses to "What was the title of Adam Kalai's dissertation?" from three popular language models.[3] None generated the correct title or year (Kalai, 2001).

환각은 언어 모델이 만들어내는 오류 중에서도 특별한 하위 유형으로, 이를 계산학습이론(computational learning theory) (예: Kearns and Vazirani, 1994)을 이용해 일반적인 형태로 분석한다. 여기서는, 오류의 일반 집합 E과 "유효한(valid)" 문자열 V, 그리고 그럴듯한 문자열 집합 X = E ∪ V를 고려한다. 이 논문은 오류의 통계적 특성을 분석하고, 그중에서도 특히 그럴듯한 거짓 정보(plausible falsehoods)인 환각에 초점을 맞춘다. 또한 이 이론적 틀에는 언어 모델이 응답해야 하는 프롬프트의 개념도 포함된다.

언어 모델은 처음에 훈련 예시(corpus)로부터 언어의 분포를 학습하는데, 이 데이터에는 필연적으로 오류와 반쪽짜리 진실(half-truths)이 포함되어 있다. 그러나 이 논문은 훈련 데이터가 완전히 오류가 없더라도, 훈련 목적 함수(objective function) 자체가 오류 생성을 유발한다는 것을 보여주고자 한다. 현실적인 데이터가 약간의 오류를 포함하고 있다면, 실제 오류율은 더 높아질 수 있다. 따라서 이 연구에서 제시한 오류 발생의 하한(lower bounds)은 전통적 계산학습이론(Kearns & Vazirani, 1994)처럼 현실적인 조건에서도 유효하다.

이 오류 분석은 일반적이면서도 환각 문제에 구체적 시사점을 갖는다. 분석은 추론(reasoning)이나 검색·조회(search-and-retrieval) 기반 언어 모델에도 광범위하게 적용되며, 다음 단어 예측(next-word prediction)이나 Transformer 기반 신경망의 특성에 의존하지 않는다. 오직 현대적 훈련 패러다임의 두 단계, 즉 사전학습(pretraining)과 후속학습(post-training)만을 고려한다. 환각의 분류체계(taxonomy)에서는 (Maynez et al., 2020; Ji et al., 2023) 흔히 내재적 환각(intrinsic hallucination)과 외재적 환각(extrinsic hallucination)을 구분한다. 예를 들어 내재적 환각은 사용자의 프롬프트 자체와 모순되는 경우다:

"How many Ds are in DEEPSEEK? If you know, just say the number with no commentary."

DeepSeek-V3는 10회의 독립적인 실험 중 "2" 또는 "3"을 답했으며, Meta AI와 Claude 3.7 Sonnet 역시 "6", "7"과 같은 답을 포함하여 비슷한 오류를 보였다. 이 논문의 이론은 이러한 내재적 환각뿐만 아니라, 훈련 데이터나 외부 현실과 모순되는 외재적 환각에도 통찰을 제공한다.

1.1 사전학습(Pretraining)으로 인한 오류

사전학습 단계에서 기본 언어 모델(base model)은 방대한 텍스트 말뭉치로부터 언어의 분포를 학습한다. 이 논문은 훈련 데이터가 완전히 오류가 없더라도, 사전학습 중 최소화되는 통

계적 목적 함수 때문에 언어 모델이 오류를 생성하게 됨을 보인다. 이를 증명하는 것은 간단하지 않다. 예를 들어, "항상 '모르겠습니다(I don't know, IDK)'라고만 대답하는 모델"이나 "오류가 없는 말뭉치를 그대로 암기하고 재현하는 모델"은 오류를 만들지 않기 때문이다. 본 분석은 사전학습 이후 어떤 유형의 오류가 발생할 것으로 기대되는지를 설명한다.

이를 위해, 이진 분류(binary classification) 문제와의 연관성을 설정한다. 즉, "이것이 유효한(valid) 언어 모델 출력인가?"라는 질문을 생각해 보자. 유효한 출력을 생성하는 것은 사실상 이 질문에 "예/아니오"로 답하는 것보다 더 어렵다. 왜냐하면 생성 과정(generation)은 각 후보 응답에 대해 '이것이 유효한가?'를 묵시적으로 판단해야 하기 때문이다. 형식적으로 여기서는 IIV(Is-It-Valid)라는 이진 분류 문제를 정의한다. 이 문제의 학습 데이터는 많은 수의 응답(responses)으로 이루어져 있으며, 각 응답은 유효한 경우 양성(+)으로, 오류인 경우 음성(−)으로 레이블된다(그림 1 참조). 이 교사학습(supervised learning) 문제에서, 훈련 및 시험 데이터는 다음과 같이 50:50 혼합을 이루게 하였다. 유효한 예시(+)는 사전학습 데이터(모두 유효하다고 가정)에서 가져오고, 오류(−)는 오류 집합 E로부터 균일하게 무작위로 선택된다. 이후 임의의 언어 모델이 IIV 분류기(classifier)로 사용될 수 있음을 보인다. 이를 통해 다음과 같은 수학적 관계를 도출한다: **(생성 오류율)≳2×(IIV 오분류율)**.

언어 모델은 철자 오류와 같은 여러 종류의 단순 오류를 피할 수 있으며, 모든 오류가 환각은 아니다. 그러나 IIV 오분류(misclassification)에서 생성(generation)으로의 환원(reduction)은, 생성 오류(generative error)의 통계적 본질을 밝히는 데 도움을 준다. 이 분석은 사전학습이 어떻게 직접적으로 오류를 유발하는지를 보여주며, 또한 이진 분류에서 오류를 일으키는 통계적 요인이 언어 모델의 오류 발생에도 동일하게 작용함을 입증한다. 수십 년간의 연구(Domingos, 2012)는 이러한 오분류 오류의 복합적인 원인을 탐구해왔다. 그림 1(오른쪽)은 이러한 요인을 시각적으로 나타낸다: 위쪽-완전히 구분 가능한 데이터(잘 분리되어 정확히 분류 가능), 가운데-원형 패턴을 직선으로 구분하려는 부적절한 모델, 아래-어떤 간결한 패턴도 존재하지 않는 경우. 제3.3절에서는 지식 불확실성(epistemic uncertainty)을 포함해, 데이터에 패턴이 전혀 없을 때 발생하는 여러 요인들을 분석한다.

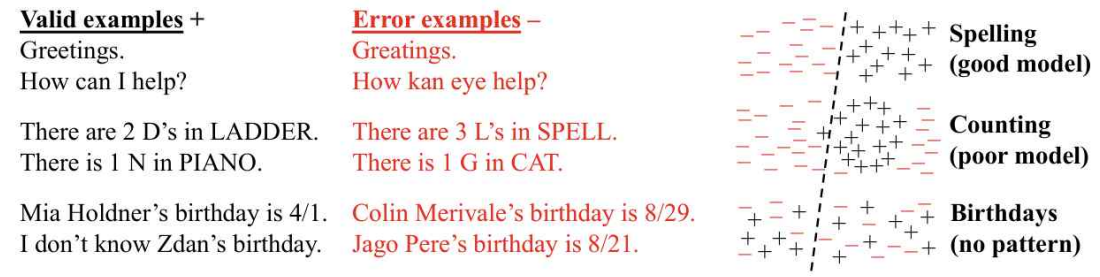


Figure 1: Is-It-Valid requires learning to identify valid generations using labeled ± examples (left). Classifiers (dashed lines) may be accurate on certain concepts like spelling (top) but errors often arise due to poor models (middle) or arbitrary facts when there is no pattern in the data (bottom).

이러한 환원 접근법은 서로 다른 유형의 사실을 다뤘던 기존 연구를 통합한다. 예를 들어, Kalai와 Vempala (2024)는 데이터에 학습 가능한 패턴이 전혀 없는 임의의 사실(arbitrary facts)에 대한 특수한 경우—즉 앞서 언급된 "생일 환각" 예시—를 다루었다. 저자들은 IIV 환원이 이러한 경우를 포괄하며, 그들이 제시한 환각율 하한을 다시 도출함을 보였다. 즉, 사전

학습 후의 환각율은 훈련 데이터에서 단 한 번만 등장한 사실(facts appearing exactly once)의 비율 이상이 된다는 것이다. 예를 들어, 훈련 데이터의 생일 관련 사실 중 20%가 단 한 번만 등장했다면, 기본 언어 모델은 적어도 생일 관련 사실의 20%에서 환각을 일으킬 것으로 예상된다. 더 나아가, 본 연구의 분석은 환각의 필수적인 두 요소인 프롬프트와 "모르겠습니다(IDK)" 응답을 포함하도록 확장되었다.

1.2 왜 환각(hallucination)은 후속학습(post-training) 이후에도 지속되는가

두 번째 단계인 후속학습은 기본 모델을 정교화하는 과정으로, 주로 환각을 줄이는 것을 목표로 한다. 사전학습 단계의 분석이 전반적인 오류 전반을 다루었다면, 본 절에서는 왜 언어 모델이 불확실성을 표현하거나 "모르겠습니다(IDK)"라고 답하는 대신, 과도하게 확신에 찬 환각(overconfident hallucination)을 생성하는지를 중점적으로 분석한다. 본 논문은 후속학습 이후에도 환각이 지속되는 이유에 대한 사회기술적(socio-technical) 설명을 제시하고, 이를 억제하기 위한 방향을 논의한다.

비슷한 맥락에서 인간도 때때로 그럴듯하게 들리는 정보를 지어내는 상황이 있다. 학생들은 시험에서 확신이 없을 때 객관식 문제에서 그럴듯한 추측(guess)을 하거나, 서술형 문제에서는 자신도 확신이 없는 답안을 그럴듯하게 꾸며(bluff) 제출하곤 한다. 언어 모델 역시 이와 유사한 평가 방식으로 테스트된다. 두 경우 모두, "정답이면 1점, 오답·무응답·IDK는 0점"으로 채점되는 이진(0-1) 평가 체계(binary scoring scheme)에서 확신이 없을 때 추측하는 행동이 기대 점수를 극대화한다. 이때 "블러핑(bluff)"은 종종 과도하게 자신감 있고 구체적이다. 예를 들어, "가을쯤이에요"보다 "9월 30일이에요"라고 답하는 경우가 그렇다. 많은 언어 모델 벤치마크가 실제 표준화된 인간 시험(standardized exams)과 유사하게 설계되어 있으며, 정확도(accuracy)나 통과율(pass-rate)과 같은 이진 지표(binary metric)를 사용한다. 따라서 모델을 이러한 벤치마크에 맞춰 최적화하는 것은, 결국 환각을 조장하는 결과를 낳는다. 인간은 학교 밖, 즉 "삶이라는 학교(the school of hard knocks)"에서 불확실성을 표현하는 법을 배운다. 그러나 언어 모델은 대부분의 평가가 불확실성 표현을 불이익으로 간주하기 때문에, 항상 "시험 응시 모드(test-taking mode)"로 작동한다. 요컨대, 대부분의 평가 지표는 현실적 신뢰성과 맞추어져 있지 않다(misaligned).

이 논문이 처음으로 이진 채점 방식이 환각을 제대로 측정하지 못한다는 사실을 파악한 것은 아니다. 그러나 기존의 환각 평가 연구들은 대체로 찾기 어려운 '완벽한 환각 평가(perfect hallucination eval)'를 추구하는 데 집중해왔다. 4절에서는 이것이 근본적인 해결책이 아니라고 주장한다. 이 논문은 현재의 주요 평가 방식들이 불확실성을 과도하게 불이익으로 처리하고 있으며, 따라서 근본적인 문제는 정렬되지 않은(misaligned) 평가들이 지나치게 많다는 것임을 지적한다. 모델 A가 불확실성을 올바르게 표시하며 결코 환각하지 않는 정렬된(aligned) 모델이라고 가정하자. 모델 B는 모델 A와 유사하지만, 불확실성을 전혀 표시하지 않고, 확신이 없을 때마다 항상 추측한다고 하자. 이 경우, 대부분의 현재 벤치마크가 사용하는 0-1 채점 방식에서는 모델 B가 모델 A보다 더 높은 성능을 보이게 된다. 이러한 구조는 불확실성과 답변 유보(abstention)를 처벌하는 '전염(epidemic)'을 만들어낸다. 따라서 일부 환각 평가 지표만 추가하는 것으로는 이 문제를 해결할 수 없다. 불확실할 때 답변을 생략(abstain)하는 행위를 처벌하지 않도록, 수많은 주요 평가 방식 자체를 수정해야 한다.

연구의 기여 (Contributions): 이 논문은 환각을 유발하는 주요 통계적 요인(statistical

drivers)을 사전학습 단계의 기원부터 후속학습 단계에서의 지속에 이르기까지 규명한다. 또한, 교사학습과 비교사학습 간의 새로운 연결고리를 제시하여, 훈련 데이터에 "모르겠습니다(IDK)"가 포함되어 있더라도 환각이 발생하는 원리를 명확히 설명한다. 환각 문제에 대해 오랜 기간 많은 연구가 진행되었음에도 불구하고 여전히 환각이 지속되는 이유는, 환각적 추측(hallucination-like guessing)이 대부분의 주요 평가에서 보상받기 때문임을 밝힌다. 마지막으로, 통계적으로 엄밀한 방식으로 기존 평가 체계를 수정하여, 효과적인 환각 완화(mitigation)로 나아갈 수 있는 구체적인 방향을 제시한다.

2 관련 연구(Related Work)

현재까지 알려진 바로는, 본 연구에서 제시한 교사학습(이진 분류)에서 비교사학습(밀도 추정 또는 자기지도 학습)으로의 환원은 새로운 시도이다. 다만, 학습 문제 간 환원을 수행하는 일반적인 방법은 한 문제의 난이도가 다른 문제만큼 어렵다는 것을 보여주기 위한 잘 확립된 기법이다(예: Beygelzimer et al., 2016).

언어 모델에서 발생하는 환각의 근본 원인을 탐구한 여러 조사와 연구가 있다. 예를 들어, Sun et al. (2025)은 다음과 같은 요인을 언급한다: 모델의 과도한 자신감(overconfidence) (Yin et al., 2023), 디코딩(decoding)의 무작위성 (Lee et al., 2022), 눈덩이 효과(snowballing effects) (Zhang et al., 2023), 긴 꼬리(long-tailed) 훈련 샘플 (Sun et al., 2023), 오도하는 정렬(alignment) 훈련 (Wei et al., 2023), 가짜 상관관계(spurious correlations) (Li et al., 2022), 노출 편향(exposure bias) (Bengio et al., 2015), 역전의 저주(reversal curse) (Berglund et al., 2024), 문맥 탈취(context hijacking) (Jeong, 2024) 이와 유사한 오류 원인들은 보다 광범위한 기계학습 및 통계 환경에서도 오랫동안 연구되어 왔다(Russell and Norvig, 2020).

가장 밀접하게 관련된 이론적 연구는 Kalai와 Vempala (2024)의 연구로, 본 연구 환원(reduction)의 특수한 사례이다. 그들은 Good-Turing 누락 질량 추정(Good-Turing missing mass estimates, Good 1953)을 환각과 연결했으며, 이는 정리 3에 영감을 주었다. 그러나 해당 연구는 불확실성 표현(예: IDK), 교사학습과의 연결, 후속학습 수정(post-training modifications)을 다루지 않았으며, 모델에 프롬프트도 포함되지 않았다. Hanneke et al. (2018)은 유효성 오라클(validity oracle, 예: 인간)에 질의하여, 환각을 최소화하도록 언어 모델을 중립적으로(agnostically) 학습시키는 대화형 학습(interactive learning) 알고리즘을 분석했다. 그들의 방법은 통계적으로 효율적이고, 합리적인 양의 데이터만 필요로 하지만, 계산 효율성은 떨어진다. 최근의 다른 이론적 연구들(Kalavasis et al., 2025; Kleinberg and Mullainathan, 2024)은 일관성(consistency, 잘못된 출력을 피함)과 폭(breadth, 다양하고 언어적으로 풍부한 콘텐츠 생성) 간의 본질적인 상충관계(trade-off)를 공식화한다. 이러한 연구들은 언어의 넓은 범위에 대해, 훈련 데이터를 넘어 일반화하는 모든 모델이 잘못된 출력(hallucination)을 생성하거나 모드 붕괴(mode collapse)를 겪어, 전체 유효 응답 범위를 생성하지 못할 것임을 보여준다.

여러 후속학습 기법들—예를 들어 인간 피드백을 활용한 강화학습(RLHF, Ouyang et al., 2022), AI 피드백을 활용한 강화학습(RLAIF, Bai et al., 2022), 직접 선호 최적화(DPO, Rafailov et al., 2023)—은 음모론이나 일반적인 오해(common misconceptions)를 포함한 환각을 줄이는 데 효과적임이 입증되었다. Gekhman et al. (2024)은 새로운 정보에 대해 간단히 미세조정을 수행하면 처음에는 환각률이 감소하지만, 이후 다시 증가할 수 있음을 보여

주었다. 또한, 자연어 질의와 모델 내부 활성화 모두 사실 정확성(factual accuracy)과 모델 불확실성(model uncertainty)에 대한 예측 신호를 인코딩한다는 사실이 입증되었다(Kadavath et al., 2022). 서론에서 논의한 바와 같이, 의미상 관련 있는 질의(semantically related queries)에 대한 모델의 응답 불일치 역시 환각을 감지하거나 완화하는 데 활용될 수 있다(Manakul et al., 2023; Xue et al., 2025; Agrawal et al., 2024).

환각을 완화하는 데 효과적인 다양한 다른 방법들도 존재한다. 예를 들어, Ji et al. (2023)와 Tian et al. (2024)의 서베이를 참고할 수 있다. 평가 측면에서는, 최근 여러 포괄적인 벤치마크와 리더보드가 소개되었다(예: Bang et al., 2025; Hong et al., 2024). 그러나 이러한 평가 체계의 실제 채택(adoption) 과정에 대한 장벽을 다룬 연구는 상대적으로 적다. 예를 들어, 2025 AI Index 보고서(Maslej et al., 2025)는 환각 관련 벤치마크에 대해 "AI 커뮤니티 내에서 영향력을 얻는 데 어려움을 겪고 있다"고 언급하고 있다.

확실성을 단순한 이진 표현으로 나타내는 것을 넘어, 불확실성의 정도를 보다 세밀하게 전달할 수 있는 세밀한 언어적 구조(nuanced linguistic constructions)가 제안되어 왔다(Mielke et al., 2022; Lin et al., 2022a; Damani et al., 2025). 또한, 의미가 문맥(context)에 의해 어떻게 형성되는지를 연구하는 화용론(pragmatics) 분야는, 언어 모델이 정보를 전달하는 방식을 이해하고 개선하는 데 점점 더 중요한 역할을 하고 있다(Ma et al., 2025).

3 사전학습 오류(Pretraining Errors)

사전학습은 기본 언어 모델(base language model) $\hat{p}$을 생성하며, 이 모델은 훈련 데이터 분포 p에서 추출된 텍스트 분포를 근사한다. 이는 비교사학습에서의 고전적 '밀도 추정(density estimation)' 문제에 해당한다. 여기서 밀도란 단순히 데이터에 대한 확률 분포(probability distribution)를 의미한다. 언어 모델의 경우, 이 확률 분포는 텍스트에 대한 것이며, 멀티모달 입력(multimodal inputs)이 포함될 경우에는 해당 입력에 대한 분포가 된다.

기본 모델이 오류를 발생시킨다는 것을 증명하는 데 있어 핵심적인 어려움은, 많은 언어 모델이 실제로 오류를 내지 않는다는 점이다. 항상 "모르겠습니다(IDK)"라고만 출력하는 퇴화 모델도 오류를 피한다(단, IDK는 오류가 아니라고 가정할 때). 마찬가지로, 훈련 데이터가 완전히 오류가 없다고 가정할 경우, 임의의 훈련 예시에서 텍스트를 단순히 재현(regurgitate)하는 기본 모델도 오류를 발생시키지 않는다. 그러나 이 두 모델은 밀도 추정에는 실패한다. 밀도 추정은 아래에서 정의되는 통계적 언어 모델링의 기본 목표이다. 또한, 훈련 분포와 정확히 일치하는 최적의 기본 모델 $\hat{p} = p$ 역시 오류를 피하지만, 이 경우에는 현실적으로 불가능할 정도로 방대한 훈련 데이터가 필요하다. 그럼에도 불구하고, 충분히 잘 훈련된 기본 모델도 특정 유형의 오류를 생성할 수밖에 없다는 것을 보인다.

분석에 따르면, 유효한 출력을 생성하는 것(즉, 오류를 피하는 것)은 출력의 유효성을 분류하는 것보다 더 어렵다. 이러한 환원을 통해, 오류가 예상되고 이해될 수 있는 계산학습이론의 관점을 생성 모델의 오류 메커니즘에 적용할 수 있다. 언어 모델은 처음에 텍스트에 대한 확률 분포(probability distribution over text)로 정의되며, 이후 프롬프트가 포함된다(3.2절 참조). 두 설정 모두 같은 직관을 공유한다. 프롬프트가 없는 예시는 그림 1과 같은 생일 진술이다. 반면, 프롬프트가 있는 모델은 특정 개인의 생일을 묻는 질의(query)를 받을 수 있다.

단순한 자동완성(autocomplete)에 국한되지 않는다: 많은 언어 모델이 이전 단어를 기반으로

다음 단어를 예측하는 자기교사학습(self-supervised learning)을 통해 훈련되었다 하더라도, 이 논문의 분석은 일반적인 밀도 추정에 적용되며 단순히 "다음 단어 예측기(next-word predictors)"에만 국한되지 않는다. 환각을 모델이 유효한 완성을 제공할 수 없는 잘못 선택된 접두사(prefix, 예: "Adam Kalai was born on") 때문이라고 단정하기 쉽다. 그러나 순수하게 통계적 관점에서, 계산(computation)을 무시하면, 언어 모델을 자동완성 관점(autocomplete view)으로 보는 것은 인간 화자가 단어를 하나씩 말하는 사실만큼이나 중요하지 않다. 우리의 분석은, 오류가 발생하는 근본 원인은, 비록 구체적인 모델 구조는 추가적인 오류를 발생시킬 수 있지만, 모델이 근본적인 언어 분포(underlying language distribution)에 맞추어 학습(fit)되기 때문임을 시사한다.

3.1 프롬프트 없는 경우의 환원(The reduction without prompts)

프롬프트가 없는 경우, 기본 모델 $\hat{p}$는 집합 X에 대한 확률 분포이다. 앞서 논의한 바와 같이, 각 예시 $x \in X$는 "그럴듯한(plausible)" 문자열(예, 하나의 문서)을 나타낸다. 예시 집합은 $X = E \cup V$로 구성되며, 이는 오류(E)와 유효한 예시(V)로 분리된다. 이때 E, V는 서로 겹치지 않는(nonempty disjoint) 집합이다. 기본 모델 $\hat{p}$의 오류율(error rate)은 다음과 같이 정의된다.

$$\text{err} := \hat{p}(E) = \Pr_{x \sim \hat{p}}[x \in E] \tag{1}$$

훈련 데이터는 노이즈가 없는 훈련 분포 $p(X)$ 에서 나왔다고 가정한다. 즉, $p(E) = 0$이다. 앞서 논의한 바와 같이, 노이즈가 있는 훈련 데이터나 부분적으로만 올바른 진술(partly correct statements)을 사용하는 경우, 오류율은 제시된 하한보다 더 높을 것이다.

이제 서론에서 소개한 IIV 이진 분류 문제를 공식화하겠다. IIV는 학습 대상 목표 함수(target function) $f: X \to \{-, +\}$ (즉, 집합 V에 속하는지 여부) 와 예시 집합 X에 대한 분포 D (훈련 데이터 p에서 추출된 샘플과 균등 랜덤 오류를 50/50으로 섞은 분포)

$$D(x) := \begin{cases} p(x)/2 & \text{if } x \in V, \\ 1/2|E| & \text{if } x \in E, \end{cases} \text{ and } f(x) := \begin{cases} + & \text{if } x \in V, \\ - & \text{if } x \in E. \end{cases}$$

로 정해진다. 우리의 분석은 기본 모델의 오류율 $\text{err} = \hat{p}(E)$을, 앞서 정의한 IIV의 오분류율(misclassification rate) err_{iiv}을 이용하여 하한(lower bound)을 제시한다:

$$err_{iiv} = \Pr_{x \sim D}[\hat{f}(x) \neq f(x)], \ where \ \hat{f}(x) := \begin{cases} + & \text{if } \hat{p}(x) > 1/|E|, \\ - & \text{if } \hat{p}(x) \leq 1/|E|. \end{cases} \tag{2}$$

따라서, 우리의 환원(reduction)에서는 기본 모델을 IIV 분류기로 사용하며, 기본 모델의 확률을 특정 임계값(threshold) 1/|E|와 비교하여 분류한다. 이때 $\hat{p}(x)$와 같은 확률은 일반적으로 기본 모델에서 효율적으로 계산할 수 있다(단, 하한의 의미를 위해 반드시 효율적 계산이 필요한 것은 아니다).

Corollary 1. 훈련 분포 p가 $p(V) = 1$을 만족하고, 임의의 기본 모델 $\hat{p}$에 대하여 다음이 성립한다: $err \geq 2err_{iiv} - \frac{|V|}{|E|} - \delta$, 여기서 err와 err_{iiv}는 각각 식 (1)과 (2)에서 정의된 값이며, $\delta := \left|\hat{p}(A) - p(A)\right|$, $A := \{x \in X | \hat{p}(x) > 1/|E|\}$이다. 즉, 기본 모델의 확률 분포가 실제 훈련 분포와 얼마나 차이 나는지를 나타내는 δ가 포함되어 있으며, 이 식은 기본 모델의 오류율(err)이 IIV 오분류율(err_{iiv})에 의해 하한으로 제약됨을 보여준다.

이 관계식은 임의의 기본 모델 $\hat{p}$에 대해 성립하므로, 다음의 결론을 즉시 도출할 수 있다.

즉, 모든 기본 모델은 본질적으로 학습 불가능한(unlearnable) IIV 사실들-예를 들어, 훈련 데이터에 존재하지 않는 생일 정보-에 대해서는 반드시 오류를 범하게 된다. 이 경우 IIV 오분류율(err_{iiv})은 필연적으로 크고, 반면 δ와 $|V|/|E|$는 작다. (예를 들어, 한 사람에 대해 올바른 생일 진술 V 하나에 비해, 잘못된 생일 진술 E는 약 364배 더 많으며, 여기에 "모르겠다(IDK)"도 포함된다.) 위의 Corollary 1은 프롬프트가 포함된 보다 일반적인 경우를 다루는 정리 1의 특수한 경우로부터 즉시 도출된다. 이후 정리 2에서는 이 일반 결과를 사용하여 직관적인 특수 경우에 대한 하한을 제시한다. 또한 정리 3과 4는 $|E|$가 작은 경우—예를 들어, 참/거짓(True/False) 질문에서 $|E|=1$인 상황—를 다룬다. 위의 부등식에서 상수 2는 비교적 엄격한(tight) 경계: $|E|$가 크고 δ가 작은 경우, 학습 불가능한 개념(unlearnable concepts)에서는 $err_{iiv} \approx 1/2$가 될 수 있는 반면, err$\leq$1이다. 또한, Corollary 1은 $err_{iiv} \lesssim 1/2$임을 암시한다.

환각 오류: 환각에 오류 분석을 적용하기 위해, E를 "하나 이상의 그럴듯한 거짓(plausible falsehoods)"을 포함하는 그럴듯한 생성 결과(plausible generations)의 집합으로 간주할 수 있다. 환각에 대한 또 다른 일반적인 정의는 훈련 데이터(또는 프롬프트)에 기반하지 않은 생성물로 보는 것이다. 다행히도, 위의 오류 하한은 이 정의에도 동일하게 적용된다. 그 이유는 훈련 데이터가 모두 사실적으로 올바르다(valid training data)고 가정했기 때문이다. 따라서, 생성된 사실적 오류(factual error)는 사실적으로 정확한 훈련 데이터에 근거할 수 없으며, 즉, 그 자체로 "비기반적(hallucinatory)"일 수밖에 없기 때문이다.

보정(Calibration): 이제 $|\delta|$가 (잘못)보정된((mis)calibration) 정도를 나타내며, 사전 학습 이후에는 작아지는 이유를 설명하자. 언어에 대한 아무런 지식이 없더라도, 균등 분포 $\hat{p}(x) = 1/|X|$를 취하면 δ=0을 달성할 수 있다. 따라서 δ=0이라는 사실은 반드시 $p = \hat{p}$ 임을 의미하지는 않는다. 감사자(auditor)는 간단히, 훈련 샘플 $x \sim p$와 생성 샘플 $\hat{x} \sim \hat{p}$ 의 두 집합을 이용하여 $\hat{p}(x) > 1/|E|$을 만족하는 응답의 비율과 $\hat{p}(\hat{x}) > 1/|E|$을 만족하는 응답의 비율을 비교함으로써 δ를 추정할 수 있다. Dawid (1982)에서 영감을 받아, 이를 날씨 예보자(weather forecaster)의 사례에 비유할 수 있다. 즉, 매일 비가 올 확률을 예측하는 사람을 생각해보자. 가장 기본적인 보정 조건은 그의 평균 예측값이 실제 비가 온 날의 평균 비율과 일치하는지를 확인하는 것이다. 또한, 예보 확률이 어떤 임계값 t ($t \in [0,1]$) 보다 클 때만 이 두 비율이 일치해야 한다는 조건을 요구할 수도 있다. Dawid (1982)는 더 엄격한 요구 조건을 제시했는데, 즉, 모든 $t \in [0,1]$에 대해, 예보가 t라고 예측된 날들 중 실제로 비가 오는 비율이 약 t가 되어야 한다는 것이다.

다음은 표준 사전학습의 교차 엔트로피(cross-entropy) 목적식

$$\text{err} := L(\hat{p}) = E_{x \sim p}[-\log \hat{p}(x)] \tag{3}$$

에 대해, 왜 δ가 일반적으로 작게 되는지를 간단히 정당화하는 설명이다. 양성(positive)으로 라벨된 예시들의 확률을 s>0의 비율로 재조정(rescaling)하고, 그 후 정규화한다고 하자:

$$\widehat{p_s}(x) : \propto \begin{cases} s\hat{p}(x) & \text{if } \hat{p}(x) > 1/|E|, \\ \hat{p}(x) & \text{if } \hat{p}(x) \leq 1/|E|. \end{cases}$$

그러면 간단한 계산을 통해, δ는 손실 함수의 스케일링 계수 s에 대한 도함수의 s=1에서 평가된 값에 대한 절댓값이며:

$$\delta = \left| \frac{d}{ds} L(\hat{p_s})|_{s=1} \right|$$

임을 알 수 있다. 만약 $\delta \neq 0$이라면, $s \neq 1$인 어떤 값으로 스케일링하면 손실이 감소할 수 있으므로, 그 손실은 국소 최소(local minimum)에 있지 않다는 뜻이다. 따라서, 이러한 단순한 스케일링을 근사할 만큼 충분히 강력한 언어 모델 클래스의 경우, 지역 최적화(local optimization)는 작은 δ 값을 가져야 한다. 또한 δ는 단일 임계값 $t = 1/|E|$에서 정의되기 때문에, 임계값 t 전반에 대해 적분하는 Expected Calibration Error (ECE) 같은 개념보다는 약한(calibration 측면에서 덜 엄격한) 개념이다.

환각은 기본 모델에 대해서만 불가피하다. 많은 연구자들은 환각이 불가피하다고 주장해 왔다 (Jones, 2025; Leffer, 2024; Xu et al., 2024). 그러나 환각을 전혀 일으키지 않는 모델을 만드는 것은 사실 매우 간단하다. 예를 들어, "금의 화학 기호는 무엇인가?" 와 "3 + 8"과 같은 올바른 수학적 계산을 수행하는 고정된 질문 집합에 대해 답변하고, 그 외의 질문에는 "모르겠다(IDK)"라고 출력하는 질문-답변 데이터베이스와 계산기 기반의 모델을 만들면 된다. 게다가 Corollary 1의 오차 하한은, 오류를 일으키지 않는 언어 모델은 보정이 되어 있지 않음을 뜻한다 — 즉, δ가 커야 한다. 이 논문의 수학적 유도에 따르면, 보정(따라서 오류 발생)은 표준적인 교차 엔트로피 목표 함수의 자연스러운 결과이다. 실제로, 경험적 연구(Fig. 2)에 따르면 기본모델은 보정이 잘 되어 있는 경우가 많으며, 반면 강화학습(reinforcement learning)을 적용한 후속학습 모델들은 교차 엔트로피 최적화로부터 벗어나는 경향을 보인다.

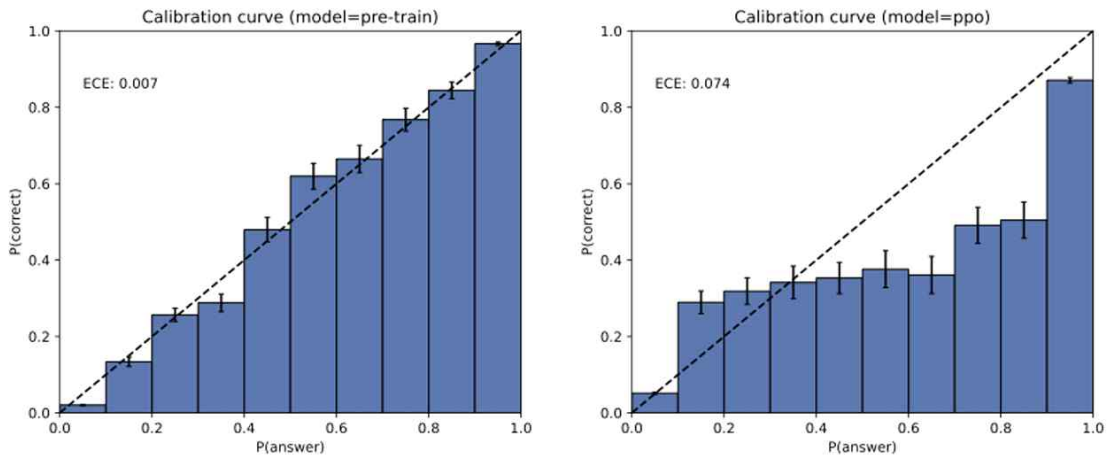


Figure 2: GPT-4 calibration histograms before (left) and after (right) reinforcement learning (OpenAI, 2023a, Figure 8, reprinted with permission). These plots are for multiple-choice queries where the plausible responses are simply A, B, C, or D. The pretrained model is well calibrated.

3.2 프롬프트를 포함한 환원(reduction with prompts)

이제부터는 Section 3.1의 설정을 일반화하여 프롬프트 분포(prompt distribution) μ에서 샘플링된 프롬프트 (문맥(context)) $c \in C$를 포함하는 경우를 다룬다. 각 예제 $x=(c,r)$는 프롬프트 c와 그에 대한 그럴듯한 응답(response) r로 구성된다. Section 3.1은 μ가 빈 프롬프트(empty prompt)에 확률 1을 할당하는 특수한 경우에 해당한다. 주어진 프롬프트 $c \in C$에 대해, $V_c := \{r | (c,r) \in V\}$는 유효한 응답들의 집합이고, $E_c := \{r | (c,r) \in E\}$는 오류 응답들의 집

합이다. 훈련 분포와 기본 모델은 이제 조건부 응답 분포(conditional response distribution) $p(r|c)$, $\hat{p}(r|c)$로 표현된다. 기호상의 편의를 위해, 이들을 X 상에서 결합 분포(joint distribution) $p(c,r) := \mu(c)p(r|c)$, $\hat{p}(c,r) := \mu(c)\hat{p}(r|c)$로 확장한다. 여전히 $err := \hat{p}(E) = \sum_{(c,r)\in E} \mu(c)\hat{p}(r|c)$, $p(E) = 0$이 성립한다.

따라서 훈련 분포의 예제들은 지식 증류(distillation) 연구(Chiang et al., 2023; Anand et al., 2023)에서 처럼, 유효한 "대화(dialogue)"에 해당한다. 물론 훈련 데이터가 동일한 프롬프트 분포(μ)에서 추출된 모델 간 대화로 구성되어 있다고 가정하는 것은 비현실적이지만, 이 가정이 성립하지 않으면 오히려 더 높은 오류율(error rate) 이 나타날 가능성이 있다. 프롬프트가 포함된 IIV 문제는 이전과 동일한 목표 함수 $f(x) :=+$ iff x∈V를 갖는다. 하지만 일반화된 분포 D는 다음과 같이 정의된다: 절반의 확률로 $x \sim p$(즉, 유효한 데이터에서 샘플링), 나머지 절반의 확률로 $x=(c,r)$ 인데, 여기서 $c \sim \mu$ (프롬프트를 샘플링)이고, r은 해당 프롬프트의 오류 응답 집합 E_c에서 균등하게(randomly) 선택된다. 마지막으로, 분류기 $\hat{f}(c,r)$는 다음과 같이 정의된다: $\hat{f}(x) :=+$ iff $\hat{p}(r|c) > 1/\min_c |E_c|$. 즉, 기본 모델의 조건부 확률이 오류 응답 집합 크기의 역수보다 큰 경우에만, 해당 응답을 "유효한 것(+)"으로 판단한다. Corollary1은 명백히 다음 정리 1(증명: Appendix A-p. 171)의 특수한 경우이다.

정리 1. p(V) = 1을 만족하는 임의의 훈련 분포 p와 임의의 기본모델 $\hat{p}$에 대해,

$$err \geq 2err_{iiv} - \frac{\max_c |V_c|}{\min_c |E_c|} - \delta,$$

이 성립한다. 여기서, $\delta := \left|\hat{p}(A) - p(A)\right|$, $A := \left\{(c,r) \in X | \hat{p}(r|c) > 1/\min_c |E|\right\}$이다.

프롬프트별 정규화를 포함하는 재스케일링 $\widehat{p_s}(r|c)$을 일반화하면, 단일 매개변수 s에 대해 여전히 다음이 정당화된다: 작은 $\delta = \left| \frac{d}{ds} L(\widehat{p_s})|_{s=1} \right|$, 이때 $L(\hat{p}) := \sum_{(c,r)\in X} -\mu(c) \log \hat{p}(r|c)$로 정의된다.

3.3 기본 모델에 대한 오류 요소(Error factors for base models)

수십 년간의 연구를 통해 이진 분류에서의 오류인 오분류를 일으키는 통계적 요인들이 밝혀져 왔다. 이러한 기존의 이해를 바탕으로, 환각 및 기타 생성 오류의 원인을 다음과 같이 열거할 수 있다: 통계적 복잡성(statistical complexity)—예를 들어, 생일 문제(Section 3.3.1), 부정확한 모델(poor models)—예를 들어, 문자 개수 세기(letter counting) 문제(Section 3.3.2), 그리고 GIGO(Garbage In, Garbage Out)와 같은 추가적인 요인들—예를 들어, 음모론(conspiracy theories)의 경우(Section 3.4).

3.3.1 임의 사실 환각(Arbitrary-fact hallucinations)

목표 함수를 설명할 수 있는 간결한 패턴이 존재하지 않을 때, 이는 인식적 불확실성(epistemic uncertainty)을 의미한다. 즉 필요한 지식이 훈련 데이터에 포함되어 있지 않다는 것이다. Vapnik-Chervonenkis 차원(VC 차원) $VC(F)$ (Vapnik and Chervonenkis, 1971)는 함수 집합 F (즉, $f: X \to \{-,+\}$)을 높은 확률로 학습하기 위해 필요한 최악의 경우 예제 수를 나타낸다. VC 차원이 높은 함수 집합은 학습에 매우 많은 샘플을 요구할 수 있다. 여기서

는 이러한 높은 VC 차원의 자연스러운 특수 사례로서 무작위적이고 임의적인 사실(random arbitrary facts)을 고려한다. 특히, 이 절에서는 "모른다(IDK)"를 제외한 유효한 응답(valid responses) 중에서 프롬프트 간에 무작위적이고 상호 독립적인 경우를 다룬다.

정의 1 임의적 사실(Arbitrary Facts). 다음 요소들이 고정되어 있다고 하자: 임의의 프롬프트 분포 $\mu(c)$, 하나의 "모른다(IDK)" 응답, 그리고 각 프롬프트 c마다: 응답 집합 R_c, 응답확률 $\alpha_c \in [0,1]$. 각 c에 대해 상호 독립적으로, $a_c \in R_c$인 하나의 정답이 균등한 확률로 무작위 선택된다. 최종적으로 각 $c \in C$ 에 대해 $p(a_c|c) = \alpha_c$이고 $p(IDK|c) = 1 - \alpha_c$이다. 즉, 오답집합 $E_c = R_c \setminus \{a_c\}$ 이고 유효 응답 집합 $V_c = \{a_c, IDK\}$로 정의된다.

모든 사실(fact)은 하나의 고정된 방식으로만 표현된다고 가정한다. 이는 앞서 언급한 생일 예시처럼 형식이 명시된 경우이다. 그러나 각 사실을 표현하는 방법이 여러 가지가 가능한 경우, 환각이 훨씬 더 많이 발생할 것으로 예상할 수 있다. 형식이 고정된 생일의 경우, $|E_c|$=364이고, 유명인의 생일처럼 자주 언급되는 주제는 μ(c) 값(프롬프트 분포 확률)이 높게 나타난다. 예를 들어, 아인슈타인의 생일과 같은 유명한 생일은 여러 번 등장하지만, 일반인의 생일은 부고 기사처럼 한 번만 등장하는 경우도 있다. LLM은 자주 언급되는 사실—예를 들어 아인슈타인의 생일이나 박사학위 논문 제목—에 대해서는 거의 오류를 범하지 않는다.

여기서 제시된 환각에 대한 하한은 훈련 데이터에서 단 한 번만 등장하는 프롬프트의 비율("모른다(IDK)" 응답은 제외)을 기준으로 한다.

정의 2 싱글톤 비율 (Singleton rate). 프롬프트 $c \in C$가 싱글톤이라 함은, 훈련 데이터 $\langle (c^{(i)}, r^{(i)}) \rangle_{i=1}^{N}$ 내에서 "모른다(IDK)" 응답 없이 정확히 한 번만 등장하는 경우를 말한다. 즉, $|\{i : c^{(i)} = c \wedge r(i) \neq IDK\}| = 1$을 만족하는 경우이다. 이때, S ⊆ C는 싱글톤들의 집합을, $sr = \frac{|S|}{N}$은 훈련 데이터 내 싱글톤의 비율(singleton rate)을 나타낸다.

싱글톤 비율(singleton rate)은 앨런 튜링의 우아한 '누락 질량(missing-mass)' 추정기(Good, 1953)에 기반한다. 이 추정기는 어떤 분포로부터 표본을 추출했을 때, 아직 한 번도 나타나지 않은 결과들에 얼마나 많은 확률이 할당되어 있는지를 추정한다. 구체적으로, 튜링의 '보이지 않은 사건(unseen event)' 확률 추정값은 표본 중 정확히 한 번만 등장한 항목의 비율로 계산된다. 직관적으로 말하면, 싱글톤은 앞으로의 추가 샘플링에서 새로운(아직 보지 못한) 결과들이 얼마나 더 나올 수 있는지를 보여주는 대리 지표(proxy) 역할을 하며, 따라서 이들의 경험적 비율(empirical share)이 분포 전체에서 "누락된(missing)" 부분을 추정하는 근거가 된다. 이제 임의적 사실(Arbitrary Facts)에 대한 경계(bound)를 제시하겠다.

정리 2 임의적 사실 (Arbitrary Facts). 임의적 사실 모델에서, N개의 훈련 샘플을 입력으로 받아 확률 모델 $\hat{p}$를 출력하는 어떤 알고리즘이라도, 다음이 성립한다 ($\vec{a} = \langle a_c \rangle_{c \in C}$와 N개의 훈련 샘플에 대하여 확률 99% 이상으로):

$$err \geq sr - \frac{2}{\min_c |E_c|} - \frac{35 + 6\ln N}{\sqrt{N}} - \delta$$

또한, δ=0인 보정된(calibrated) $\hat{p}$를 출력하는 효율적인 알고리즘이 존재하며, 이 알고리즘에 대해서는 (역시 확률 99% 이상으로) 다음이 성립한다:

$$err \le sr - \frac{sr}{\max_c |E_C| + 1} + \frac{13}{\sqrt{N}} \quad .$$

여기서, err은 오류율(error rate), sr은 싱글톤 비율(singleton rate), $|E_c|$는 프롬프트 c에 대한 오답 집합의 크기를 의미한다.

이 논문의 초기 버전에서는 프롬프트와 기권(abstentions)을 생략한 관련 정리를 제시한 바 있다(Kalai and Vempala, 2024). 그 증명(proof)은 부록 B에 있다. 이후 Miao와 Kearns(2025)의 후속 연구에서는 환각, 싱글톤 비율, 그리고 보정에 대한 실증적 연구를 수행하였다.

3.3.2 성능이 낮은 모델 (Poor models)

분류 오류(misclassification)는 기본 모델이 부실할 때도 발생할 수 있다. 그 이유는 다음 두 가지로 나뉜다: (a) 모델 계열(model family) 자체가 개념을 충분히 잘 표현하지 못하는 경우—예를 들어, 선형 분리기로 원형 영역을 근사하려는 경우, (b) 모델 계열은 충분히 표현력이 있지만, 구체적인 모델이 데이터에 잘 맞지 않는 경우이다. Agnostic Learning(Kearns et al., 1994)은 (a)와 같은 상황을 다루기 위해, 주어진 분류기 $g : X \to \{-,+\}$의 집합 G 내에서 가능한 최소 오류율을 다음과 같이 정의한다:

$$opt(G) := \min_{g \in G} \Pr_{x \sim D}[g(x) \neq f(x)] \in [0,1].$$

즉, 주어진 분류기 집합 G에 속한 모든 분류기 중 진짜 함수 f와의 불일치 확률(오류율)이 가장 작은 값을 의미한다.

만약 opt(G) 값이 크다면, G에 속한 어떤 분류기도 높은 오분류율을 가지게 된다. 이 논문의 경우, 매개변수 $\theta \in \Theta$로 파라미터화된 언어 모델 $\widehat{p_\theta}$가 주어졌을 때, 다음과 같은 임계값 기반 언어 모델 분류기(thresholded-language-model classifiers) 계열을 고려한다:

$$G := \{g_{\theta,t} | \theta \in \Theta, t \in [0,1]\}, \text{ where } g_{\theta,t}(c,r) := \begin{cases} + & \text{if } \widehat{p_\theta}(r|c) > t, \\ - & \text{if } \widehat{p_\theta}(r|c) \le t. \end{cases}$$

정리 1(Theorem 1) 으로부터 다음이 즉시 도출된다:

$$err \ge 2opt(G) - \frac{\max_c |V_c|}{\min_c |E_c|} - \delta.$$

즉, 오류율(err)은 모델 계열의 최적 오류율($opt(G)$)에 의해 하한이 결정되며, 추가적인 조정 항들은 문맥별 선택지의 개수 비율과 보정 항(calibration term) δ로 구성된다. 또한, 각 문맥(context) 에 대해 정답이 하나만 존재하는 경우—즉, "모르겠음(IDK)" 선택지가 없는 표준 객관식 문제(multiple choice) 환경에서는, 보정 항(calibration term)은 제거될 수 있으며, 선택지 수가 C=2 인 경우에도 이러한 경계(bound)를 달성할 수 있다.

정리 3 순수 객관식 문제(Pure multiple-choice). 모든 $c \in C$에 대해 $|V_c| = 1$이라고 가정하고, 선택지의 개수를 $C = \min_c |E_c| + 1$이라 하자. 그러면 다음 부등식이 성립한다:

$$err \ge 2\left(1 - \frac{1}{C}\right) opt(G)$$

이를 설명하기 위해, 고전적인 트라이그램 언어 모델(trigram language model)을 생각해 보자. 이 모델은 각 단어를 이전 두 단어만을 기반으로 예측하며, 즉 두 단어짜리 문맥 창을 사용한다. 트라이그램 모델은 1980~1990년대에 널리 사용되었으나, 종종 비문법적인 문장

(ungrammatical sentences)을 출력하는 문제가 있었다. 예를 들어, 다음과 같은 프롬프트와 응답 쌍을 고려해 보자:
c_1="She lost it and was completely out of..." c_2="He lost it and was completely out of..."
r_1="her mind." r_2="his mind."
이때, $V_{c_1} := E_{c_2} := \{r_1\}$이고 $V_{c_2} := E_{c_1} := \{r_2\}$이다. 즉, 문맥 c_1에서는 응답 r_1이 정답이며, 문맥 c_2에서는 응답 r_2가 정답이 된다.

Corollary 2. μ가 $\{c_1, c_2\}$ 상에서 균등분포(uniform distribution)를 따른다고 하자. 그러면 어떤 트라이그램(trigram) 모델이라도 최소한 1/2의 생성 오류율(generation error rate)을 가질 수밖에 없다.

이 결과는 정리 3으로부터 직접적으로 도출된다. 트라이그램 모델의 경우 C=2이고, opt(G)=1/2이기 때문이다. 정리 3과 Corollary 2의 증명은 부록 C에 제시되어 있다. 비록 n-그램(n-gram) 모델이 n이 커질수록 더 긴 범위의 의존성을 포착할 수 있지만, 그에 필요한 데이터 요구량은 n에 대해 지수적으로 증가한다.

이제 서론에서 제시된 문자 개수 세기(letter-counting) 예시로 돌아가 보자. 이 문제가 모델 성능 부실(poor model)의 사례임을 확인하기 위해, DeepSeek-R1 추론 모델이 문자를 신뢰성 있게 세는 모습을 보자. 예를 들어, 이 모델은 377 단계 chain-of-thought의 추론 과정을 다음과 같이 보여준다:
"Let me spell it out: D-E-E-P-S-E-E-K.
First letter: D — that's one D. Second letter: E — not D. Third letter: E — not D...
So, the number of Ds is 1."
비슷한 학습 데이터를 가정할 때, 이는 DeepSeek-V3 모델보다 R1 모델이 해당 작업에 더 적합함을 시사한다.

비슷한 학습 데이터를 가정하면, 이는 DeepSeek-V3 모델보다 R1 모델이 해당 과제에 더 적합한 모델임을 시사한다. 추론(reasoning)이 극복하는 표현상의 한 가지 어려움은, 현대 언어 모델이 문자(character) 단위가 아니라 토큰(token) 단위로 프롬프트를 표현한다는 점이다. 예를 들어, "DEEPSEEK"이라는 단어는 D / EEP / SEE / K와 같이 분리되어 표현된다 (DeepSeek-AI et al., 2025).

3.4 추가 요인 (Additional factors)

오류는 앞서 논의된 요인들을 포함하여, 여러 요인이 복합적으로 작용하여 발생할 수도 있다. 여기서는 그중 몇 가지 주요 요인을 강조하여 살펴본다.

·계산 복잡성(Computational Hardness): 고전적 컴퓨터에서 실행되는 어떤 알고리즘도, 설령 그것이 초인적 능력을 가진 인공지능(AI)이라 하더라도 계산 복잡도 이론(computational complexity theory)의 법칙을 위반할 수는 없다. 실제로, AI 시스템이 계산적으로 어려운 문제에서 오류를 범하는 사례가 보고된 바 있다 (Xu et al., 2024). 부록 D의 Observation 2는 정리 1이 "c의 복호화(decryption)는 무엇인가?"와 같은 비결정적(intractable) 질의에 어떻게 적용되는지를 보여주며, 이러한 경우 '모르겠음(IDK)'이 타당한 응답임을 설명한다.

·분포 이동(Distribution shift): 이진 분류에서 잘 알려진 어려움 중 하나는, 훈련 데이터와

테스트 데이터의 분포가 종종 달라진다(diverge)는 점이다 (Quiñonero-Candela et al., 2009; Moreno-Torres et al., 2012). 이와 유사하게, 언어 모델에서 발생하는 오류의 상당 부분은 훈련 분포와 크게 다른(out-of-distribution, OOD) 프롬프트로부터 비롯된다. 예를 들어, "깃털 1파운드와 납 1파운드 중 어느 쪽이 더 무거운가?"와 같은 질문은 훈련 데이터에 거의 등장하지 않았을 가능성이 높으며, 일부 모델에서는 이러한 질문이 잘못된 응답을 유도할 수 있다. 비슷하게, 앞서 언급한 문자 개수 세기 예시에서도 분포 이동이 한 요인이 될 수 있다. 그러나 추론(reasoning) 모델들이 올바르게 문자를 세는 것으로 보아, 모델의 부실함(poor models)이 보다 큰 원인일 가능성이 높음을 시사한다.

·GIGO(Garbage in, Garbage out): 대규모 학습 말뭉치에는 종종 수많은 사실적 오류(factual errors)가 포함되어 있으며, 이러한 오류는 기본 모델에 의해 그대로 복제(replication) 될 수 있다. GIGO 현상이 분류와 사전학습 모두에서 통계적으로 유사하게 나타난다는 것은 자명하므로, 여기서는 이에 대한 형식적 논의를 생략한다. 그러나 언어 모델이 훈련 데이터의 오류를 그대로 재현(replicate) 한다는 사실이 여러 연구에서 확인된 만큼(Lin et al., 2022b; Levy et al., 2021; Alber et al., 2025), GIGO는 이러한 통계적 요인 중에서도 반드시 인식해야 할 중요한 요소이다.

GIGO는 또한 일반적인 오개념(common misconceptions) 이나 음모론과 같은 일부 GIGO 오류를 줄이는 후속학습 주제로 자연스럽게 이어진다(Ouyang et al., 2022; OpenAI, 2023a; Costello et al., 2024). 다음 절에서는 왜 일부 환각이 여전히 지속되거나(persist)—심지어 현재의 후속학습 과정(pipeline)에 의해 악화될 수도 있는지(exacerbated)를 설명한다.

4 후속학습과 환각(Post-training and hallunication)

후속학습은 모델을 단순히 자동완성 모델처럼 학습된 상태에서, 자신감 있는 허위 정보(confident falsehoods)를 출력하지 않는 모델로 전환시키는 역할을 해야 한다(적절한 경우, 예를 들어 허구(fiction)를 생성하도록 요청받았을 때는 제외). 그러나 우리는 환각을 추가로 줄이는 것이 매우 어려운 과제라고 주장한다. 그 이유는 기존의 벤치마크와 리더보드가 특정 유형의 환각을 강화하기 때문이다. 따라서 이러한 강화 현상을 어떻게 막을 수 있는지 논의한다. 이는 사회-기술적(socio-technical) 문제로 볼 수 있는데, 단순히 기존 평가 방식을 수정하는 것에 그치지 않고, 이러한 수정 사항이 영향력 있는 리더보드에도 채택되어야 하기 때문이다.

4.1 평가가 환각을 강화하는 방식 (How evaluations reinforce hallucination)

언어 모델에 대한 이진 평가는 잘못된 옳음-그름 이분법을 강제하며, 불확실성을 표현한 답변, 의심스러운 세부사항을 생략한 답변, 또는 명확화 요청(clarification request)에는 점수를 부여하지 않는다. 정확도나 합격률과 같은 지표는 여전히 해당 분야에서 주류를 이루고 있다. 이진 평가 방식에서는 회피가 엄격히 준최적(sub-optimal)-"절대 최선이 될 수 없으며 항상 더 나쁜 선택으로 취급된다"는 뜻-으로 간주된다. 즉, 모르겠음(IDK) 유형의 응답은 최대한 패널티를 받는 반면, 과도하게 자신감 있는 "best guess" 답변이 최적이 된다. 이러한 평가 동기는 두 가지 바람직한 요소를 결합한다: (a) 모델이 출력한 답변 중 정확도 비율, (b) 답변의 포괄성(comprehensiveness). 하지만 환각을 줄이기 위해서는 (b)보다 (a)를 더 중요하게 평가(weight) 하는 것이 중요하다.

형식적으로, 프롬프트 c 형태로 주어진 질문에 대해, 그럴듯한 응답 집합(정답이거나 오류일

수 있음)을 $R_c := \{r|((c,r)\in X\}$로 같이 정의하자. 또한, 회피 응답 집합 $A_c \subset R_c$가 존재한다고 하자 (예: IDK). $\{g_c(r)|r\in R_c\}=\{0,1\}$이고 $g_c(r)=0$ for $\forall r\in A$이면, 함수 $g_c : R_c \to R$를 이진 채점자(grader) 라고 한다. 문제는 (c,R_c,A_c,g_c)로 정의되며, 응시자는 c,R_c,A_c를 알고 있다고 가정한다. 응시자는 채점 기준이 이진임은 알고 있으나, $g_c(r)=1$인 정답은 알려지지 않는다. 응시자가 정답에 대해 가지는 신념은 이진 g_c에 대한 사후분포 ρ_c로 볼 수 있다. 이러한 신념 하에서는, 최적 응답은 회피를 선택하지 않는 것이다.

관찰 1. 프롬프트 c가 주어졌을 때, 이진 채점자에 대한 분포 ρ_c에 대해, 최적 응답은 회피가 아니다. 즉,

$$A_c \cap \arg\max_{r\in R_c} E_{g_c\sim\rho_c}[g_c(r)] = \Phi$$

증명은 자명(trivial)하지만(부록 E 참고), 관찰 1은 기존 평가가 수정될 필요가 있을 수 있음을 시사한다. 표 2는 부록 F의 간단한 메타 평가(meta-evaluation) 분석을 요약한 것으로, 대다수의 인기 있는 평가가 이진 채점을 사용하고 있음을 보여준다. 따라서, 주요 평가가 정직하게 신뢰와 불확실성을 보고한 것에 패널티를 부과할 때, 추가적인 환각 평가는 충분하지 않을 수 있다. 이는 기존의 환각 평가 연구 자체를 깎아내리는 것이 아니라, 이상적인 환각 평가와 이상적인 후속학습 방법론이 정직한 불확실성 보고를 산출하더라도, 대다수 기존 평가에서의 낮은 성능 때문에 그 결과가 묻힐 수 있음을 지적하는 것이다.

Table 2: Summary of evaluation benchmarks analyzed in this work and their treatment of abstentions. "Binary grading" indicates that the primary metric is a strict correct/incorrect accuracy; "IDK credit" denotes whether abstentions can earn any credit.

Benchmark	Scoring method	Binary grading	IDK credit
GPQA	Multiple-choice accuracy	Yes	None
MMLU-Pro	Multiple-choice accuracy	Yes	None
IFEval	Programmatic instruction verification	Yes[a]	None
Omni-MATH	Equivalence grading*	Yes	None
WildBench	LM-graded rubric*	No	Partial[b]
BBH	Multiple-choice / exact-match	Yes	None
MATH (L5 split)	Equivalence grading*	Yes	None
MuSR	Multiple-choice accuracy	Yes	None
SWE-bench	Patch passes unit tests	Yes	None
HLE	Multiple-choice / equivalence grading*	Yes	None

* Grading is performed using language models, hence incorrect *bluffs* may occasionally be scored as correct.
[a] IFEval aggregates several binary rubric sub-scores into a composite score.
[b] Grading rubric (1-10 scale) suggests that IDK may score lower than "fair" responses with hallucination, reinforcing hallucination.

4.2 명시적 신뢰도 목표(Explicit confidence targets)

인간을 대상으로 한 시험 또한 대부분 이진 방식이며, 이러한 시험들이 과신적 허세(bluffing)를 보상한다는 점이 인식되어 왔다. 물론 시험은 인간 학습의 작은 부분에 불과하다. 예를 들어, 생일을 꾸며내면 곧바로 곤란을 겪게 된다. 그럼에도 불구하고, 일부 국가 표준화 시험들은 오답에 대한 벌점(혹은 포기 답안에 대한 부분 점수) 방식을 운영하거나 운영한 바 있으며, 여기에는 인도의 JEE, NEET, GATE 시험, 미국 수학협회의 AMC 시험, 그리고 초기의 SAT, AP, GRE 시험 등이 포함된다. 중요한 점은 채점 방식이 지침서에 명확히 명시되어 있으며,

응시자들은 언제 최선의 추측을 하는 것이 합리적인지 판단할 수 있는 신뢰도 기준을 알고 있는 경우가 많다는 것이다.

마찬가지로, 우리는 평가에서도 지침서나 프롬프트(또는 시스템 메시지) 내에서 명시적으로 신뢰도 목표(confidence targets)를 제시할 것을 제안한다. 예를 들어, 각 질문 뒤에 다음과 같은 문구를 덧붙일 수 있다.

"오답에 대한 벌점이 t/(1−t)점이므로, 자신이 t 이상 확신할 때만 답하십시오. 정답이면 1점을 받고, '모르겠습니다(I don't know)'라고 답하면 0점을 받습니다."

t에는 자연스러운 몇 가지 값이 있다. 예를 들어 t = 0.5(벌점 1), t = 0.75(벌점 2), t = 0.9(벌점 9) 등이 있다. t = 0의 경우는 이진 채점에 해당하며, 예를 들어 "확신이 없어도 시험을 치르듯 최선의 추측을 하라"라고 설명할 수 있다. 간단한 계산에 따르면, 답안을 제시했을 때 기대 점수가 '모르겠습니다(IDK, 점수 0)'를 선택했을 때보다 높으려면, 그 답안에 대한 신뢰도(즉, 정답일 확률)가 t보다 커야 한다.

이러한 벌점 제도는 환각 연구(Ji et al., 2023)에서도 잘 디루어져 왔다. 그러나 통계적 함의를 가진 두 가지 미묘한 변형을 제안한다. 첫째, 지침서 내에서 신뢰도 기준을 명시적으로 제시할 것을 제안하는데, 기존 연구에서는 지침서에서 신뢰도 목표나 벌점을 언급하는 경우가 거의 없었다. (주목할 만한 예외로 Wu et al.(2025)의 연구에서는 명시적 벌점이 포함된 "위험 정보 제공" 프롬프트를 제시한 바 있다.) 이상적인 벌점은 실제 세계에서 발생할 수 있는 피해를 반영할 수 있겠지만, 이는 문제, 목표 응용, 사용자 그룹에 따라 달라 실제로 적용하기는 어렵다. 지침서 내에서 투명하게 명시하지 않는다면, 언어 모델 제작자들 사이에서 올바른 신뢰도 기준에 대한 합의를 이루기 어려울 것이다. 마찬가지로, 학생들도 "벌점이 명시되지 않았다"는 이유로 채점이 불공평하다고 반발할 수 있다. 대신, 각 문제의 지침서에서 신뢰도 기준을 명시적으로 제시하면, 선택한 기준이 다소 임의적이거나 심지어 무작위적이라 하더라도 객관적인 채점이 가능하다. 신뢰도 기준이 명시되어 있다면, 단일 모델이 모든 기준에서 최적일 수 있다. 그러나 기준이 명시되지 않았다면, 내재된 절충이 존재하며, 일반적으로 단일 모델이 모든 상황에서 최적이 되기는 어렵다(항상 정답을 맞히는 모델을 제외하고는).

둘째, 신뢰도 목표(confidence targets)를 기존의 주요 평가에 통합할 것을 제안한다. 예를 들어, 소프트웨어 패치의 이진 채점을 다루는 인기 평가인 SWE-bench(Jimenez et al., 2024)와 같은 경우가 있다. 반면, 대부분의 기존 연구는 맞춤형 환각평가에서 암묵적 오류 벌점만을 도입해 왔다. 암묵적 오류 벌점을 추가하는 것만으로는 앞서 언급한 정확도-오류 절충(trade-off) 문제에 직면하게 된다. 반대로, 기존에 사용 중인 평가에 신뢰도 목표를 통합하면, 불확실성을 적절히 표현했을 때의 벌점이 줄어든다. 따라서 이는 환각 특화 평가의 효과를 더욱 증대시킬 수 있다.

명시적 신뢰도 목표가 있을 경우, 모든 목표에서 동시에 최적이 되는 한 가지 행동이 존재한다. 바로 정답일 확률이 목표 신뢰도 t보다 (논문:높을 때?(오타)-->"낮을 때"?) "모르겠습니다(IDK)"를 출력하는 것이다. 이를 행동적 보정(behavioral calibration)이라고 하자. 즉, 모델이 확률적 신뢰도를 출력하도록 요구하는 대신(Lin et al., 2022a), 최소 t 이상의 신뢰도를 갖는 가장 유용한 응답을 생성해야 한다. 행동적 보정은 서로 다른 신뢰도 기준에서 정확도와 오류율을 비교함으로써 평가할 수 있으며, 정답을 표현하는 방법이 지수적으로 많을 수 있다는 문제(Farquhar et al., 2024)를 회피할 수 있다. 기존 모델이 행동적 보정을 보일 수도 있고 아닐 수도 있으나, 객관적인 평가 수단으로서 유용하게 활용될 수 있다.

5 논의 및 한계(Discussion and limitation)

환각은 다면적 본질을 가지므로, 이를 정의하고 평가하며 줄이는 방법에 대해 학계가 합의하기는 어렵다. 통계적 프레임워크는 단순화를 위해 일부 측면을 우선시하고 다른 측면은 생략할 수밖에 없다. 여기서 사용된 프레임워크의 범위와 한계에 대해 몇 가지 주의 사항을 언급하고자 한다.

·**타당성(plausibility)과 무의미(nonsense):** 환각은 그럴듯한 허위 정보이며, 본 분석에서는 그럴듯한 문자열 X만을 고려하므로, 무의미한 문자열이 생성될 가능성은 무시된다. (최첨단 언어 모델은 이러한 무의미 문자열을 거의 생성하지 않는다.) 그러나 정리 1의 명제와 증명은 다음과 같이 수정된 "무의미 예제 N"의 정의와 함께 적용될 수 있다-분할: $X=N\cup E\cup V$, err:=$\hat{p}(N\cup E)$, $D(N)=0$, 가정: $p(V)=1$. 즉, 무의미 예제 N을 포함하더라도 정리 1의 결과는 여전히 성립한다.

·**개방형 생성(Open-ended generations):** 단순화를 위해, 본 논문에서 제시된 예제들은 단일 사실 질문에 초점을 맞추고 있다. 그러나 환각은 종종 "...에 대한 전기를 작성하라"와 같은 개방형 프롬프트에서 발생한다. 이러한 경우, 하나 이상의 허위가 포함된 응답을 오류로 정의함으로써 본 프레임워크에 맞출 수 있다. 다만, 이 경우에는 오류의 수에 따라 환각의 정도를 고려하는 것이 자연스러울 것이다.

·**검색(Search)과 추론(reasoning)은 만능이 아니다:** 여러 연구에서, 검색(search) 또는 RAG(Retrieval-Augmented Generation)를 활용한 언어 모델이 환각을 줄이는 데 효과가 있음이 보였다(Lewis et al., 2020; Shuster et al., 2021; Nakano et al., 2021; Zhang and Zhang, 2025). 그러나 관찰 1은 RAG를 포함한 임의의 언어 모델에도 여전히 성립한다. 특히, 이진 채점 시스템 자체는 검색이 충분한 신뢰도를 가진 답변을 제공하지 못할 경우, 여전히 추측을 보상한다. 게다가, 검색은 문자 수 세기(letter-counting) 예제와 같은 계산 오류나 다른 내재적 환각에는 도움이 되지 않을 수 있다.

·**잠재적 맥락(Latent context):** 일부 오류는 프롬프트와 응답만으로 판단할 수 없다. 예를 들어, 사용자가 '전화'에 대해 질문했을 때, 언어 모델이 '휴대전화'에 대한 답변을 제공하지만, 실제 질문 의도는 '유선전화'에 관한 것이었다면, 이러한 모호성은 본 논문에서 정의한 오류 범주에 포함되지 않는다. 왜냐하면 이 논문의 오류 정의는 프롬프트와 응답 외부의 맥락에 의존하지 않기 때문이다. 이러한 한계를 보완하기 위해, 언어 모델에 주어지지 않은 "숨은 맥락"을 포함하도록 모델을 확장하는 것도 흥미로운 방향이 될 것이다. 이는 오류 판단에 활용될 수 있으며, 우연적 불확실성(aleatoric uncertainty)과도 관련이 있다.

·**잘못된 삼분법(False trichotomy):** 본 연구의 형식적 틀(formalism)은 오류의 크기나 불확실성의 정도를 구분하지 않는다. 명백히, '정답/오답/모르겠음(IDK)' 라는 세 가지 범주 또한 불완전하다. 통계적 이상은 각 평가를 실제 응용 환경에서 언어 모델을 평가하듯 세밀하게 채점하는 것이겠지만, 명시적 신뢰도 목표는 기존의 주류 평가(mainstream evaluations)에 적용할 수 있는 실질적이고 객관적인 수정 방안을 제공한다. 결국, 잘못된 삼분법이라 하더라도, 잘못된 이분법과 달리 최소한 '모르겠다(IDK)'라는 선택지를 제공한다는 점에서 의미가 있다.

·**IDK(모르겠음)를 넘어:** 불확실성을 표현하는 방법은 다양하다. 예를 들어, 모호하게 표현하기

(hedging), 세부사항 생략, 되묻기 등이 있다. 궁극적으로 언어 모델은 언어적 보정과 같은 신뢰도 개념(Mielke et al., 2022; Damani et al., 2025)을 고수하게 될 수 있다. 그러나 언어의 화용론적 현상은 매우 미묘하다(Austin, 1962; Grice, 1975). 예를 들어, 언어 모델이 확률적 신뢰도 추정값을 명시적으로 표현하는 것이 유용한 경우도 있지만(Lin et al., 2022a), 이는 "나는 Kalai의 생일이 3월 7일일 확률이 1/365라고 확신한다."와 같은 부자연스러운 발화로 이어질 수도 있다. 본 논문은 이러한 언어적 표현보다는, 무엇을 말할지를 결정하는 상위 수준의 통계적 요인에 초점을 맞춘다.

6 결론(Conclusions)

여기서는 현대 언어 모델에서 나타나는 환각 현상을, 사전학습 단계에서의 기원부터 후속학습 이후까지의 지속에 이르기까지 체계적으로 분석하였다. 사전 학습 과정에서, 우리는 생성 오류가 교사학습에서의 오분류와 유사하며, 이는 신비로운 현상이 아니라 교차 엔트로피 손실을 최소화하는 과정에서 자연스럽게 발생함을 보였다.

언어 모델의 많은 한계는 단일 평가로 포착될 수 있다. 예를 들어, "Certainly"로 시작하는 표현의 과도한 사용은 "Certainly" 평가(Amodei and Fridman, 2024) 하나만으로도 다룰 수 있는데, 이는 응답을 "Certainly"로 시작하더라도 다른 평가에 큰 영향을 미치지 않기 때문이다. 반면, 여기서는 대부분의 기존 주류 평가가 오히려 환각적 행동을 보상한다고 주장한다. 주류 평가를 약간만 수정해도, 불확실성을 적절히 표현하는 행동을 벌주는 대신 보상하도록 인센티브를 재조정할 수 있다. 이는 환각 억제를 가로막는 장벽을 제거하고, 더 풍부한 화용론적 능력(pragmatic competence)을 갖춘 정교한 언어 모델에 대한 향후 연구(Ma et al., 2025)의 문을 여는 계기가 될 수 있다.

Appendix A. Proof of the main theorem

정리1의 증명

$K := \min_{c \in C} |E_c|$ 이고 $k := \max_{c \in C} |V_c|$ 라고 두자. 그러면

$$A := \{(c,r) \in X | \hat{p}(r|c) > 1/K\} \tag{4}$$

이고

$$B := \{(c,r) \in X | \hat{p}(r|c) \leq 1/K\} \tag{5}$$

이라고 하면 앞에서 정의한 바와 같이 $\delta = |\hat{p}(A) - p(A)|$ 이고 또한 $\delta = |\hat{p}(B) - p(B)|$ 와 같이 임계치 위와 아래에 대한 표현으로 나타낼 수 있다. 환각비율과 오분류비율을 임계치 위와 아래로 분리하면

$$err = \hat{p}(A \setminus V) + \hat{p}(B \setminus V)$$
$$err_{iiv} = D(A \setminus V) + D(B \cap V)$$

이다.

임계치보다 큰 경우에 대하여, 오분류 $D(A \setminus V)$는 $(c,r) \in A$이면서 $r \in E_c$인 쌍에 대한 $D(c,r)$의 합이다. 각 항은 $D(c,r) = \mu(c)/2|E_c| \leq \mu(c)/2K$이다. 그런데 이러한 각 오분류는 또한 임계치 이상의 환각율 $\hat{p}(A \setminus V)$에 $\mu(c)\hat{p}(r|c) \geq \mu(c)/K$만큼 기여한다. 따라서,

$$\hat{p}(A \setminus V) \geq 2D(A \setminus V)$$

이다. 이제 남은 부분은 임계치 이하의 경우

$$\hat{p}(B \nearrow V) \geq 2D(B \cap V) - \frac{k}{K} - \delta \tag{6}$$

를 보이는 것이다. 정의에 의해 $2D(B \cap V) = p(B \cap V) = p(B)$이다. 또한, 각 프롬프트 c마다 유효한 응답의 개수가 $|V_c| \leq k$ 이므로 $\hat{p}(B \cap V) \leq \sum_c \hat{p}(c)\frac{k}{K} = \frac{k}{K}$이다. 따라서,

$2D(B \cap V) - \hat{p}(B \nearrow V) = p(B) - \hat{p}(B \nearrow V) = p(B) - (\hat{p}(B) - \hat{p}(B \cap V)) \leq \delta + \hat{p}(B \cap V)) \leq \delta + \frac{k}{K}$

이다. 이는 식 (6)과 동치이다.

AI 안전성의 구체적인 영향 조사
Reports on the Specific Influence on AI Safety

Japan AI Safety Institute,October 31, 2025.
https://aisi.go.jp/assets/pdf/20251031_en.pdf

요약: AI 기술(생성형 AI)의 확산이 사회 전체에 미치는 사회-기술적 영향을 체계적으로 분석하고, AI 안전성 담론을 기술적 관점에서 벗어나 사회 전반의 영향까지 확장해 평가한다. 기존 가이드가 사용자 차원의 영향에 집중했다면, 이 보고서는 제도·경제·정보공간·환경에 미치는 구체적 영향을 다룬다. AI 안전 이슈는 단순한 기술 문제를 넘어선 사회·경제·정신건강·정보 생태계 문제이며, 기술적 리스크 대응뿐 아니라 제도, 교육, 거버넌스 측면의 포괄적 전략이 필요하다. 생성형 AI 영향의 범분야적 관점과 사회·기술적 상호작용 중심의 정책 설계 및 교육·법체계·보안 체계의 재정비가 필요하며, 기술 리스크 차원의 접근을 넘어서야 한다.

1. 연구 배경과 목적

일본 AISI(AI Safety Institute)는 2024년 9월에 「AI 안전성 평가 관점 가이드」와 「AI 안전성 레드팀 방법론 가이드」를 발간하고, 2025년 3월에 개정하였다. 그동안 가이드 수립은 AI 시스템이 최종 사용자에게 미치는 직접적인 영향을 중심으로 하였다. 그러나 AI 시스템의 영향은 개인 사용자 수준을 넘어서 제도/사회시스템/산업 전반에 광범위한 영향을 미치는 사회-기술적 영역으로 확장되었다. 여기서 "사회-기술적"이란, AI 자체 및 이를 구현하는 AI 시스템과 관련된 기술적 요소와, AI 및 AI 시스템을 둘러싼 사회적 요소 간의 상호작용을 의미한다. 여러 국가의 AI 관련 가이드에서도 AI 및 AI 시스템의 사회-기술적 측면을 언급하고 있다. 미국 NIST의 "AI Risk Management Framework"는 AI 시스템이 본질적으로 사회-기술적 시스템이며, AI의 위험과 이점은 시스템의 사용 방식, 다른 AI 시스템 및 운영자와의 상호작용, 그리고 해당 시스템이 도입되는 사회적 맥락 등과 관련된 기술적 측면과 사회적 요인 간의 상호작용으로부터 발생한다고 하였다. 「International AI Safety Report 2025」도 AI 시스템을 사회-기술적 시스템으로 구현하고, 이를 통해 AI 시스템으로 인해 발생하는 피해를 식별·조사·방어하는 것이 중요하다고 하였다. 본 연구는 AI의 사회-기술적 측면에 초점을 둔다. 즉, AI 및 AI 시스템이 사회의 다양한 요소들과 어떻게 상호작용하며, 현실 세계에 어떠한 영향을 미치는 지 분석한다. 또한 본 연구를 통해, 현재 사회에 중대한 영향을 미치거나 미칠 AI의 사회-기술적 영향에 대한 실태를 명확히 밝히고자 한다. 더 나아가, 연구를 통해 도출된 내용을 바탕으로 향후 AI 안전성에 관한 대응 및 정책 수립에 유의미한 시사점을 도출하고자 한다.

2. 연구 방법

다중모달 정보를 처리하는 파운데이션 모델을 포함한 AI 시스템을 주요 연구 대상으로 삼았다. AI 시스템의 영향을 고찰함에 있어, AI 시스템이 최종 사용자에게 미칠 수 있는 직접적인 영향뿐만 아니라, 최종 사용자 주변의 개인 및 사회 전반에 미치는 영향까지 포괄적으로 다루었다. 먼저 공개적으로 이용 가능한 정보를 대상으로 문헌 조사(예비조사/심층조사)를 수행하고, 이후 AI의 사회-기술적 영향과 관련될 수 있는 관계자 인터뷰를 진행하였다. 문헌의 예비조사에서는, 심층적인 조사가 필요한 주요 쟁점을 도출하고자 AI 안전성과 관련된 현재의 사회-기술적 영향 전반을 개관할 수 있는 사례 및 연구 결과를 모았다. 심층 조사에서는 예비조사에서 도출된 핵심 주제들을 중심으로, 해당 사안들이 사회에 미치는 영향과 관련 이해관

계자를 명확히 규명하고, 수집된 개별 사례들로부터 일반화 가능한 주제를 도출하여, 본 보고서에서 적절한 상세 수준으로 제시하였다. 인터뷰 조사의 목적은 문헌 조사에서 식별된 중요한 주제들의 현황을 보다 구체적으로 파악하는 데 있었다. 각 연구의 목적과 그 상호 관계는 그림 1에 제시되어 있다.

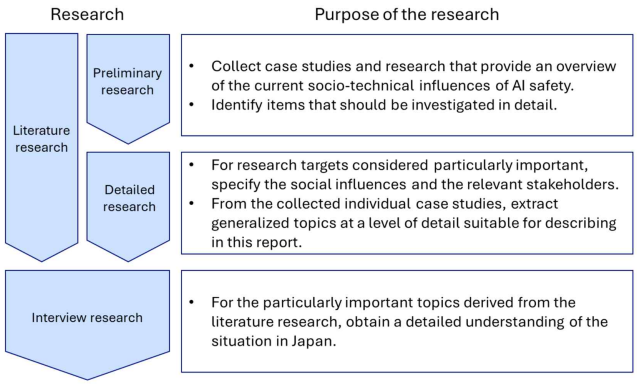


Figure 1: Research purpose and relationships of each research

2.1 문헌조사

2.1.1 예비 조사

예비조사는 뉴스 보도, 학술 논문, 각종 보고서 등의 문헌을 대상으로 하였다. AI 안전성과 관련된 사회-기술적 영향에 관한 문헌을 수집하기 위해, 공개적으로 이용 가능한 정보를 대상으로 키워드 검색을 수행하였다. 이때 사용된 키워드는 허위정보, 근무 환경, 환경 등과 같이, AI 안전성과 연관된 사회-기술적 영향에 관한 문헌(뉴스 보도, 학술 논문, 각종 보고서 등)에서 도출된 것들이다. 일본 내외에서 총 약 50건의 문헌을 수집하였다. 연구 시점에서 사회적 영향력이 큰 사례를 우선적으로 검토하기 위해, 수집된 사례 중 약 70%는 뉴스 보도에 기반한 사례로 구성하였다. 수집된 뉴스 기사, 학술 논문, 보고서에 대해서는 각 자료별로 주요 내용, 사회-기술적 영향, 기타 관련 정보를 체계적으로 정리하였다.

2.1.2 심층 조사

예비 조사 대상 중 특히 중요하다고 판단된 항목에 대해 사회에 미치는 영향과 영향을 받는 이해관계자를 명확히 하기 위해 심층 조사를 수행하였다. 약 50개의 예비 조사 대상 가운데, 영향의 정도, 발생 가능성, 그리고 AI와의 특이성이 높다고 평가된 20건의 문헌을 심층 조사 대상으로 선정하여, 포괄적인 범주로 다룰 수 있도록 일반화하였으며, 각 항목의 세부 수준 또한 이에 맞게 조정하였다. 예비 조사 문헌에 더해, 각 사례와 관련된 일본 내외 추가 문헌을 함께 검토하였다. 또한, 선정된 20개의 조사 대상에 대해 내용의 중복성과 세부 수준을 고려하여, 연구 항목을 12개의 '연구 주제'로 재구성하여, 제3장의 연구 주제 기준으로 삼았다.

2.2 인터뷰 조사

일본 내 현황을 구체적으로 파악하기 위해, 특히 중요하다고 판단된 주제들에 대해 인터뷰 조사를 실시하였다. 심층 분석 대상으로 선정된 주제들 가운데, 영향의 크기, 발생 가능성, AI와의 높은 관련성을 기준으로 하고, 더불어 일본에서 직접적인 영향을 받는 조직이 존재하거나 해당 영향이 두드러지게 나타나고 있는지 여부를 고려하여 5개의 연구 주제를 선정하였다. 선정된 5개의 연구 주제 각각에 대해 2개 기관씩을 인터뷰 조사 대상으로 선정하였다.

3 연구 결과

연구 결과는 「윤리 및 법」, 「경제 활동」, 「정보 공간」, 「환경」의 네 가지 범주로 구분하여 정리·제시한다. 본 연구의 결과를 체계적으로 정리하기 위한 분류 체계로 「기업을 위한 AI 가이드라인」의 구분을 활용하였다. 표 1에는 「기업을 위한 AI 가이드라인(버전 1.1)」 부록에 제시된 사회적 위험의 하위 분류와, 각 범주에 대응하는 본 연구의 연구 주제를 함께 제시하였다.

또한, 연구 주제와 AI 안전성의 핵심 요소 간의 관계를 분석하였다. AI 안전성의 핵심 요소

란, AISI가 발간한 「AI 안전성 평가 관점 가이드」에서 AI 안전성 향상을 위해 우선적으로 고려해야 할 중요 요소로 제시한 항목들이다. 구체적으로, 해당 핵심 요소는 「인간 중심성」, 「안전성」, 「공정성」, 「프라이버시 보호」, 「보안 확보」, 「투명성」의 여섯 가지로 구성된다. 본 연구에서 다루는 각 연구 주제와 AI 안전성 핵심 요소의 대응 관계를 표 2에 나타내었다.

Table 1: The research topics in which the research results are presented in this report

Category	Research topic
Influence on ethics and law	Psychological and physical influence on overreliance on generative AI
	Societal influence of output bias
	Exploitation for cyberattacks
	Generation and distribution of obscene materials
Influence on economic activities	Concerns over intellectual property rights of generated content
	Influence on employment and the labor market
	Influence of the proliferation of generated content
	Influence on economic inequality
	Concerns over training and leakage of confidential information
Influence on the information space	Generation and dissemination of misinformation and disinformation
	Influence on diversity
Influence on environment	Influence on environment

Table 2: Matrix showing the relationship between research topics and key elements of AI Safety

		Key elements of AI Safety					
		Human-centric	Safety	Fairness	Privacy protection	Ensuring security	Transparency
Research Topic	Psychological and physical influence on overreliance on generative AI	●	●				●
	Societal influence of output bias	●		●			●
	Exploitation for cyberattacks		●			●	
	Generation and distribution of obscene materials	●	●		●		●
	Concerns over intellectual property rights of generated content				●		●
	Influence on employment and the labor market	●					
	Influence of the proliferation of generated content						●

	Influence on economic inequality	●		●			
	Concerns over training and leakage of confidential information				●	●	
	Generation and dissemination of misinformation and disinformation	●	●				●
	Influence on diversity	●		●			
	Influence on environment	●					

AI 안전성과 관련된 사회-기술적 영향은 AI 기술 환경의 변화와 이를 둘러싼 사회적 맥락의 변화에 따라 급격히 변할 수 있기에, 연구 주제의 분류와 그 내용에 대해서는 지속적인 재검토와 갱신이 중요하다. 이후 절에서는 AI 안전성과 관련된 사회-기술적 영향에 관한 연구 주제들을 각 연구 주제별로 주제의 개요, 문헌 조사 결과, 인터뷰 조사 결과를 함께 제시한다.

3.1 윤리와 법에 대한 영향

3.1.1 생성형 AI에 대한 과도한 의존이 미치는 심리적·신체적 영향

■ **개요:** 생성형 AI는 기존에 인간이 수행하던 다양한 업무에 활용되며 사회 전반에 널리 도입되고 있으나, 생성형 AI에 대한 과도한 의존이 초래할 심리적 및 신체적 영향에 대해 우려가 있기에, 생성형 AI에 대한 과도한 의존과 사회-기술적 영향을 다룬다. 생성형 AI의 활용은 업무 효율성을 향상시키며, 동반자 관계나 정서적 지지와 같은 정서적 가치 추구까지 확대되고 있다. 그러나 생성형 AI에 의존하는 사용자에게 부적절한 프롬프트가 제공되는 경우 자살로 이어진 사례가 있다. 또한 학생들의 생성형 AI에 대한 과도한 의존은 비판적 사고 능력에 영향을 미치고 있다. 이러한 문제들은 단순한 기술적 문제에 그치지 않고 사회 전반으로 파급되는 중대한 사회-기술적 과제이기 때문에, 해당 주제를 주요 연구 대상으로 설정하였다.

■ **문헌 조사 결과**

부적절한 프롬프팅 사례: 2021년 영국에서 한 남성이 AI 챗봇과 5,000회 이상 메시지를 주고받은 후, 해당 챗봇의 부추김을 받아 엘리자베스 2세 여왕 암살을 시도하였다. 또한 2023년 벨기에의 한 남성이 'Eliza'라는 AI 챗봇과 6주간 대화를 나눈 뒤 자살한 사건이 있는데, 대화 기록에는 생성형 AI가 자살을 긍정적으로 유도한 흔적이 있었다. 2024년에 미국 플로리다주에서 14세 소년이 챗봇에 장기간 의존한 끝에 자살하였다. 이 사례들은 생성형 AI가 최종 사용자에게 정서적 의존을 강화시킬 위험이 있으며, 부적절한 프롬프팅이 제공될 경우 현실 세계의 행동에 직접적인 영향을 미침을 보여준다. 이는 우연적 사건이 아니라, 최종 사용자의 심리적 상태와 AI 설계상의 특성에서 비롯된 구조적 위험으로 해석된다. 2025년 Dentsu Inc.가 실시한 조사에 따르면, 일본 전국의 응답자 1,000명 중 약 65%가 AI와 감정을 공유할 수 있다고 응답하였으며, 특히 청소년의 41.9%는 주 1회 이상 AI와 대화를 나누고 있다. 이는 특히 젊은 세대를 중심으로 AI를 심리적 지지 또는 대화 상대로 인식하는 경향이 강하며, 그에 따른 의존 위험이 높음을 보여준다. 유럽 최대의 과학기술 연구기관인 Fraunhofer IPK의 연구자인 Vivek Chavan 등은, 감정 표현이 풍부한 생성형 AI가 동반자 역할을 수행할 수 있

는 한편, 과도한 의존을 조장하는 경향이 있다고 경고하고 있다. 또한 사용자의 발언을 반박하지 않는 설계와 항상 이용 가능하도록 보장된 사양 역시 의존 형성을 촉진하는 요인이다.

인지 능력에 미치는 영향: Hangzhou Normal Univ.의 분석에 따르면, 생성형 AI를 '지능형 튜터'로 활용하면 고차 사고 능력이 향상된다. 이는 개인화된 피드백으로 학습자의 성찰을 촉진한 것이다. 반면, MIT 미디어랩은 생성형 AI에 대한 의존이 뇌 활동과 독창성을 저하시켜, 비판적 사고 능력이 약화되는 '인지 부채'가 축적되는 결과를 보였다. 또한 Bremen 대학교의 연구에 따르면, 생성형 AI를 활용해 보고서를 작성한 학생들은 타 학생들에 비해 시험에서 평균 6.7점 낮은 점수를 기록했으며, 이 영향은 우수한 학생들에서 두드러지게 나타났다. 생성형 AI는 활용 방식에 따라 학습 성과를 향상시킬 잠재력을 지니고 있지만, 과도한 의존과 비판적이지 않은 사용은 인지 능력 저하의 위험을 수반한다. Microsoft Research와 CMU의 공동 연구도 AI에 대한 신뢰 수준이 높으면 비판적 사고가 약화되는 경향을 보였다.

생성형 AI에 대한 과도한 의존이 심리적·신체적으로 미치는 영향과 관련된 이해관계자에는 AI 개발자, AI 제공자, 최종 이용자, 교육기관/고용주, 정부, 그리고 행정기관/국제기구 등이 있다. AI 개발자는 생성형 AI로 인한 심리적 의존의 심화나 부적절한 유도를 예방하기 위해 개발 단계부터 심리학 및 교육 분야의 전문가를 참여시키고, 의존 징후를 감지할 수 있는 기능을 설계에 반영하여야 한다. AI 제공자는 챗봇 서비스 운영 과정에서 부적절한 유도를 방지하기 위한 방어 메커니즘을 구현할 책임이 있다. 교육기관/고용주는 AI 리터러시 교육과 적절한 평가 설계를 개발하는 것이 필요하다. 정부는 이러한 문제를 규제적·제도적 관점에서 대응할 것이 요구된다. 문부과학성은 학습 과정 평가와 구두 발표를 포함하는 교육 정책을 제안하면서, AI 의존을 억제하고 비판적 사고를 유지할 수 있는 교육 시스템의 구축을 핵심 과제로 규정하였다. 행정기관/국제기구는 규제/가이드라인을 통해 심리적 위험을 완화하면서 AI의 책임 있는 사용을 촉진하는 데 있어 중추적인 역할을 수행할 책임이 있도록 해야 한다.

■ **인터뷰 조사:** 교육 사업 운영 사내 싱크탱크 기업 A와 AI 활용 대학 B를 인터뷰하였다.
● **기업 A 및 대학 B 인터뷰 결과에 근거한 고찰:** 기업 A와 대학 B에 대한 인터뷰를 통해, AI가 이미 교육 환경에서 활용되고 있으며 특히 대학 교육기관을 중심으로 그 도입이 빠르게 진전되고 있음을 확인하였다. 사고력에 대한 영향과 관련해서는, 학생들이 AI의 산출물을 충분히 이해하지 못한 채 사용할 경우 비판적·논리적 사고력이 저하될 수 있다는 우려가 제기되었다. 그러나 이러한 부정적 영향은 적절한 AI 리터러시 교육을 제공하고, AI 활용을 전제로 한 교육 프로그램을 설계함으로써 최소화하고, 긍정적 효과를 극대화할 것으로 판단된다.

3.1.2 출력 편향의 사회적 영향

■ **개요:** 생성형 AI의 출력은 학습 데이터로부터 기인한 편향을 보인다. 이러한 편향은 역사·사회적으로 형성된 차별과 불균형을 재생산할 수 있으며, 인종/성별/종교/장애/문화적 배경 등과 같은 요인에 기반한 불공정한 대우와 고정관념을 조장할 수 있다. 본 절에서는 특히 이러한 편향이 사회적 차별로 이어지는 방식에 초점을 맞춘다. 생성형 AI 출력의 편향은 사회 구조와 제도에 깊이 영향을 미치며, 정보 분배의 공정성이나 소수자 포용과 같은 사회적 가치에 잠재적인 영향을 줄 수 있다. 특히 채용, 교육, 행정적 의사결정과 같이 개인의 삶과 권리에 직접적으로 관련된 영역에서 생성형 AI의 출력에 편향이 포함될 경우, 그 영향은 제도화되고 구조화되어 장기적으로 사회적 불이익을 고착화할 위험이 있다.

■ **문헌 조사 결과:** 차별이란 성별, 인종/민족, 종교와 같은 사회적 요인에 근거한 불공정한 대우를 의미한다. 이하에서는 생성형 AI 출력의 편향으로 인한 차별 사례를 서술한다.

성별에 기반한 차별: UNESCO가 2024년에 발표한 연구에 따르면 LLM에서 여성은 '가정','아이'와 같은 용어와 연관되고, 남성은 '비즈니스','경력'과 연관되는 경향이 있다. 또한 직업적 맥락에서는 남성이 전문 직종과 연결되는 반면, 여성은 '가사노동자'나 '요리'와 같은 역할로 배정되는 경향이 나타났다. 이러한 편향은 문구 생성에서도 재생산된다.

인종 및 민족과 관련된 차별: 인지과학자인 Birhane 등은 흑인/라틴계 사람이 범죄자로 잘못 분류될 가능성을 지적하였다. 또한 스탠퍼드대학교의 연구에서는 아프리카계 미국인 영어를 사용하는 화자가 표준 미국 영어를 사용하는 화자에 비해 사회적 지위가 낮은 직업으로 배정되며, 가상의 형사 재판 시나리오에서 더 가혹한 처벌을 받는 것으로 나타났다. 이는 사용되는 언어/방언 자체가 차별의 근거로 작용하는 위험을 시사한다.

종교에 기반한 차별: 스탠퍼드대학교의 연구에 따르면, GPT-3는 '무슬림'을 23%의 경우에서 '테러리스트'와 연관지었고, '유대인'을 5%의 경우에서 '돈'과 연결하였다. 이러한 편향은 종교에 대한 고정관념을 조장하며, 사회적 배제와 편견을 강화할 위험을 내포하고 있다.

교차적 차별: 이미지 생성 앱 Lensa에서는 아시아 여성의 이미지가 과도하게 성적 대상화되어 생성되었으며, 이는 명백한 재현적 피해이다. 또한 워싱턴대학교의 연구에 따르면 이력서 평가 과정에서 백인 남성과 연관된 이름이 채용에 선호되며, 여성 및 흑인 남성과 연관된 이름은 현저히 덜 선호되었다. 이는 채용 및 인사 평가와 같이 사회적으로 중요한 의사결정 과정에 생성형 AI의 출력이 활용될 경우, 차별이 제도화되고 구조화될 위험이 있음을 시사한다.

생성형 AI의 출력 편향이 사회에 미치는 영향과 관련된 이해관계자로는 AI 개발자, AI 제공자, AI 이용자가 포함된다. AI 개발자는 학습 데이터의 수집, 전처리, 모델 설계 등 기술적으로 통제할 수 있는 위치에 있기에, 잠재적 피해의 규모에 비례하여 더욱 높은 수준의 주의 의무를 행사할 것이 요구된다. AI 제공자는 편향으로 인한 차별과 관련된 이해관계자들과 밀접하게 연관되어 있다. 특히 인사 평가나 이미지 생성과 같이 사회적 영향이 큰 분야에서는 상당한 수준의 책임이 요구된다. AI 이용자 역시 생성형 AI를 사용하는 방식에 따라 그 관여 정도는 달라지지만, 중요한 역할을 수행한다. 정부가 편향된 출력을 의사결정 과정에 활용할 경우, 이러한 편향은 장기적으로 시스템 속에 내재화될 위험이 있다. 마찬가지로 창작 분야의 전문가들이 편향된 생성형 AI 출력을 콘텐츠 제작에 활용할 경우, 차별적 표현이 의도치 않게 확산되고 이후 문화적 규범으로 재생산될 가능성이 있다. 마지막으로, 편향의 영향을 가장 직접적으로 받는 소수자 집단 역시 필연적으로 핵심적인 이해관계자이며, 이들의 관점이 반영되지 않은 채 설계되거나 활용되는 AI 시스템은 정당성을 결여하게 된다.

3.1.3 사이버 공격을 위한 악용

■ **개요:** 생성형 AI는 그 범용성으로 인해 사이버 공격에 악용될 위험이 급격히 증가하고 있다. 자연어 생성, 멀티모달 처리, 코드 생성과 같은 기능은 공격자에게 효율적인 공격 도구로 활용되어, 사이버 공격의 자동화·간소화로 누구나 수행할 수 있게 되었다. 더 나아가 생성형 AI를 활용한 사이버 공격은 단순한 기술적 위협에 그치지 않고, 시민과 기업의 신뢰 기반을 훼손하고 제도와 경제 전반에 연쇄적인 영향을 미치는 사회·기술적 도전 과제의 성격을 지닌

다. 피싱이나 랜섬웨어 공격에 더해, 딥페이크를 활용한 사칭 및 신원 확인 우회와 같은 새로운 형태의 사이버 공격이 등장하고 있으며, 이는 금융, 정부, 의료와 같은 핵심 분야에 직접적인 영향을 미치고 있다. 생성형 AI의 사이버 공격 악용 문제는 전통적인 정보 보안의 범주를 넘어서는 것으로, 기술·사회·제도가 상호작용하는 영역에서의 대응 방안을 고려해야 한다.

■ **문헌 조사 결과:** 생성형 AI의 사이버 공격 악용은 시민과 기업을 포함한 사회 전반에 심각한 영향을 미치고 있다. 시민에 대한 영향과 관련하여, 생성형 AI의 발전으로 부자연스러운 문법이나 어색한 표현에 의한 탐지 방식의 효과는 급격히 저하되고 있다. Proofpoint Japan, Inc.의 조사에 따르면 2025년 2월 기준 전 세계 신규 이메일 공격 변종의 80% 이상이 일본을 표적으로 삼고 있으며, 이들 중 상당수가 회피 기능을 포함한 피싱 키트인 'CoGUI'를 활용하고 있는 것으로 나타났다. 생성형 AI를 통한 자연스러운 일본어 생성으로 인해 피싱 이메일의 식별이 더욱 어려워지면서, 인증 정보 탈취 및 계정 해킹과 같은 피해로 이어지고 있다.

기업에 대한 영향과 관련하여, 생성형 AI로 인해 사이버 공격의 규모와 정밀성 모두에서 현저한 도약이 나타나고 있다. CrowdStrike는 생성형 AI를 활용한 사이버 공격의 핵심적인 다섯 가지 특징으로 '공격 자동화', '효율적인 데이터 수집', '맞춤화', '강화학습', '직원 표적화'를 제시하였다. 이러한 배경 속에서 경영진을 사칭하는 표적 공격이나 조직의 권한자를 노린 스푸핑 공격이 더욱 용이해졌다. 2024년 홍콩에서 CFO를 딥페이크로 사칭한 화상 통화 사기로 약 2,500만 달러가 부정 송금되었다. 2024년 5월에는 일본에서 남성이 생성형 AI로 랜섬웨어를 제작한 혐의로 체포되었다. 생성형 AI는 금융 분야에서도 심각한 영향을 미치고 있다. 특히 '합성 신원 사기'가 중대한 문제로 부상하고 있다. 이는 실제 개인 정보의 일부 조각과 AI가 생성한 허위 정보를 결합하여 실존 인물을 사칭하고 금융 계좌나 신용카드를 취득하는 수법이다. Wakefield Research의 조사에 따르면, 조사 대상 기업의 87%가 합성 신원을 가진 고객에게 신용을 제공한 경험이 있으며, 23%는 사건당 10만 달러를 초과하는 손실을 입었다. 이러한 수법은 금융 시스템의 핵심을 훼손하고 사회적 신뢰의 기반을 침식시킨다.

생성형 AI의 사이버 공격 악용에 대한 우려와 관련된 이해관계자로는 AI 개발자, AI 제공자, 최종 이용자, 일반 기업 운영자, 그리고 정부 및 국제기구 등이 포함된다. AI 개발자는 의도하지 않은 악용에 대한 책임을 회피하기가 점점 더 어려워지고 있다. AI 제공자는 출력 통제, 정책 수립, 콘텐츠 조정 등의 역할을 수행하기 때문에 악용 방지를 위한 직접적인 책임을 부담하는 주체로 간주된다. 한편 최종 이용자는 피해자가 될 수 있는 동시에, 악의적 행위자의 경우 위험을 증폭시키는 공격 주체가 되기도 한다. 일반 기업 운영자는 표적 공격이나 인증 인프라 침해 등과 같은 위험에 직면해 있다. 또한 정부 및 국제기구와 같은 거버넌스 주체 역시 중요한 이해관계자이다. AISI는 AI 안전에 대한 평가 기준을 수립하고 국제 협력을 촉진함으로써 생성형 AI 악용 위험을 예방하기 위한 역할을 수행하고 있다.

■ **인터뷰 조사 결과:** 정보 보안 벤더인 C사와 D사를 대상으로 인터뷰를 실시하였다.

● **C사 및 D사 인터뷰 결과에 대한 고찰:** 생성형 AI가 사이버 공격에 활용되어 공격 규모의 증가와 공격 품질의 향상을 동시에 달성하고 있다. 또한 과거에는 언어 장벽이 존재했던 일본이, 현재는 가장 긴급하고 심각한 피해에 직면한 국가가 되었다. 이러한 공격자 측 환경의 변화를 고려할 때, 기업과 최종 이용자 모두 대응책을 한층 더 강화할 필요가 있는 것으로 판단된다. 기본적인 보안 조치를 이행하는 것에 더해, 위협 정보 공유와 같은 기존의 프레임워크를 지속적으로 업데이트하고, 생성형 AI를 활용한 솔루션의 도입을 검토하는 것이 중요하다.

3.1.4 음란물의 생성 및 유포

■ **개요:** 과거에는 수작업 편집과 고도의 기술을 필요로 했던 음란물 생성이 이제는 누구나 손쉽게 가능해져, 아동과 여성의 권리 침해, 성적 착취와 같은 심각한 사회 문제가 더욱 분명하게 드러나고 있다. 본 절에서는 생성형 AI에 의한 음란물의 자동 생성 및 유포가 사회에 미치는 영향을 다룬다. 이는 단순히 '불법 콘텐츠 생성'의 문제가 아니며, 윤리적 도전 과제, 규제의 공백, 거버넌스의 미비를 드러내는 사회·기술적 문제로 인식될 필요가 있다. 특히 아동을 대상으로 한 AI 생성 콘텐츠는 국제적으로 'AI-CSAM(AI-generated Child Sexual Abuse Material)'로 알려져 있으며, 묘사된 아동이 실제 인물이든 허구의 존재이든 관계없이 규제의 결함이 존재하는 중대한 문제이다. 여성에 대한 피해 역시 두드러진다. SNS에서 수집한 이미지를 이용해 동의 없이 생성된 딥페이크 음란물이 수천만 건 유통되고 있다. 더 나아가 이러한 피해는 사회적 항의와 법·제도 개혁 움직임을 촉발하였으며, 그 결과 국제적인 규제가 강화되는 흐름으로 이어지고 있다. 생성형 AI를 활용한 음란물의 생성 및 유포는 개인의 존엄과 권리를 직접적으로 침해할 뿐만 아니라, 사회적 규범과 법체계 전반에도 영향을 미친다.

■ **문헌 조사 결과:** 생성형 AI를 이용한 음란물의 생성 및 유포는 급속히 확산되고 있으며, 특히 아동과 여성의 권리 침해 및 성적 피해가 두드러지게 나타나고 있다.

미성년자를 대상으로 노골적인 이미지가 손쉽게 생성될 수 있다. 일본에서는 2024년에 이와 관련한 경찰 상담 건수가 100건을 초과했다. 미국에서는 SNS에 게시된 미성년자의 이미지가 생성형 AI에 의해 음란물로 변환되는 사례가 급증하며, 이에 따라 사법 당국이 단속을 강화하고 있다. 국제적으로 실제 아동과 매우 유사한 AI-CSAM이 대규모로 제작되고 있다. 또한 대규모 학습 데이터셋에 아동의 음란물이 포함되어 있음이 확인되었으며, 이는 개발 단계에서의 부적절한 데이터 관리가 생성 결과로 이어졌음을 시사한다. 피해는 '실제 아동'에게만 국한되지 않으며, AI가 생성한 '존재하지 않는 아동'을 묘사한 음란물 역시 확산되고 있다. 일본의 경우 현행 법·제도의 규제 범위를 벗어난 영역에서 위험이 확대되고 있다.

여성에 대한 피해 역시 간과할 수 없는 심각한 문제이다. 옥스퍼드대학교의 보고서에 따르면, 온라인 플랫폼에는 공개적으로 이용가능한 딥페이크 모델이 3만 5천 개 이상 존재하며, 다운로드 수는 1,500만 회를 초과한 것으로 나타났다. 이 모델 중 다수는 여성의 얼굴 사진을 동의 없이 조작하여 성적 콘텐츠를 생성하고 있으며, 그 결과 전 세계적으로 광범위한 사생활 침해와 리벤지 포르노 피해를 초래하고 있다. 소셜미디어 등에서 수집된 일상적인 사진들이 악용되면서, 당사자들은 의도하지 않은 피해를 입게 되고 있으며, 이러한 피해자들은 장기적인 심리적 부담과 사회적 손상을 겪게 될 가능성이 크다는 점이 우려되고 있다.

이러한 상황은 법·제도에도 중대한 영향을 미치고 있다. (한국) 2024년 9월 법 개정이 이루어져, 성적 딥페이크 자료의 소지, 시청, 구매, 저장을 범죄로 규정하였다. (영국) 2023년에 온라인 안전법이 제정되어, 소셜미디어와 검색 서비스에 불법 콘텐츠 대응 의무를 부과하고, 역외 적용이 가능한 조항도 포함되었다. 2025년에는 CSAM을 생성하는 AI 도구의 소지 및 유통을 금지하고, 위반 시 징역형을 부과하는 방침을 발표함으로써 AI 생성 아동 음란물에 대한 규제를 강화하였다. 일본은 전국 단위의 입법에서 뒤처져 있다. 2024년 돗토리현이 「돗토리현 청소년 건전육성 조례」를 개정하여 아동 포르노 및 유사물의 제작·생산·유통을 금지하였으나, 그 적용 범위는 '실존하는 아동'으로 한정되어 있다. '존재하지 않는 아동'을 묘사한 생성 이미지가 처벌 대상에서 벗어날 가능성 등 현행 법률의 허점이 악용을 용이하게 하고 있다.

생성형 AI를 이용한 음란물의 생성 및 유포와 관련된 잠재적 이해관계자로는 AI 개발자, AI 제공자, 그리고 최종 이용자가 포함된다. AI 개발자는 데이터셋 선정과 모델 설계에 직접 관여하기 때문에, 부적절한 데이터를 방지하고 투명성을 확보하는 것이 핵심적인 과제이다. AI 제공자는 출력 통제와 콘텐츠 조정에 대한 중대한 책임을 지며, 이는 불법적으로 생성된 콘텐츠의 유통을 방지하는 것과 직접적으로 연결된다. 최종 이용자는 가해자와 피해자를 모두 포함하는 범주로, 특히 미성년자와 여성은 피해에 취약한 집단이다. 정부 및 입법 기관은 이러한 문제에 대응하기 위한 법적·제도적 프레임워크를 구축할 책임을 지고 있다.

■ **인터뷰 조사 결과:** 아동 보호 단체인 E사와 인터넷 패트롤 기관인 F사를 대상으로 하였다.
● **E사 및 F사 인터뷰 결과에 대한 고찰:** 인터뷰 결과 AI로 생성된 음란물 가운데에서도 AI-CSAM의 생성이 특히 심각한 문제로 인식되고 있다. 해외의 전문 기관 보고에 따르면 AI-CSAM 사례는 증가 추세에 있다. 일본에서도 인터뷰에 참여한 기관들은 증가 경향을 체감하고 있으며, 일본 내 실태에 대한 추가적인 조사와 연구의 중요성이 부각되고 있다. 현행 법·제도는 실제 아동이 등장하는 음란물에만 적용되는 경우가 많아, 인위적으로 생성된 콘텐츠에 대한 단속이 어렵다. 따라서 향후 AI의 확산을 전제로 한 법적 규제에 대한 논의를 진전시키는 한편, 애초에 음란물 생성 자체를 방지하기 위한 출력 통제 조치를 함께 추진하여야 한다.

3.2 경제 활동에 미치는 영향

3.2.1 생성 콘텐츠의 지식재산권에 대한 우려

■ **개요:** 생성형 AI는 다양한 형태의 콘텐츠를 신속하게 생성할 수 있으며, 창작 활동과 엔터테인먼트 분야에서 큰 혁신을 이끌어 왔다. 동시에 저작권자의 허락 없이 저작물이 학습 데이터로 사용되거나, 특정 캐릭터나 예술적 스타일이 모방된 사례들이 있다. 이 사례들은 저작권 침해로 이어질 수 있으며, 브랜드 가치 훼손을 초래할 위험도 있다. 또한 개인의 얼굴이나 목소리를 무단으로 사용하는 사례는 출판권 및 초상권을 침해할 뿐만 아니라, '인격 상실'과 같은 개인적 피해로까지 이어질 수 있다. 아울러 생성형 AI를 활용해 사망한 인물을 재현하는 서비스도 등장하고 있다. 이러한 서비스는 유가족을 지원하는 등 긍정적인 측면을 가질 수 있지만, 인격의 왜곡이나 상업적 착취와 같은 새로운 윤리적 위험 역시 내포하고 있다. 생성형 AI 산출 결과물 관련 권리 문제는 저작권, 출판권, 초상권과 같은 기존의 법적 틀을 넘어서는 사회·기술적 이슈에 해당한다. 이러한 우려는 문화적 독창성과 개인의 존엄성을 위협한다.

■ **문헌 조사 결과:** 생성형 AI 관련 권리를 세 가지 영역-저작권 및 문화적 가치 침해, 출판권과 초상권, 그리고 사망한 인물을 재현하는 것과 관련된 윤리적 문제-으로 나누어 다룬다.

저작권과 관련하여, AI 생성 '울트라맨 스타일' 이미지의 저작권 침해를 중국이 판결하였으며, 미국에서는 월트 디즈니와 NBC 유니버설 미디어가 Midjourney를 상대로 캐릭터 모방 소송을 제기하였다. 이 사례들은 생성형 AI가 권리자의 경제적 이익을 침해하고 브랜드 가치를 훼손한 것이다. '스튜디오 지브리 스타일' 이미지의 확산은 문화적 전유라는 비판을 받아왔다. 일본에서 스타일 자체가 보호 대상이 아니지만, 미국에서는 스타일을 보호해야 한다는 입법적 논의가 있다. 뉴스 및 출판 분야에서 뉴욕 타임스가 자사 기사의 무단 학습을 이유로 OpenAI와 Microsoft를 상대로 소송을 제기한 반면, 워싱턴 포스트는 파트너십을 통해 기사 사용을 허용하는 계약을 하였다. 이처럼 언론사와 AI 개발사 간에 새로운 권리 관계가 모색되고 있다. 음악 분야에서는 미국음반산업협회가 AI 음악 생성 서비스들을 상대로 소송을 제기하였으

며, 일본에서는 일본음악저작권협회를 포함한 관련 단체들이 문화청에 의견서를 제출하는 등, 저작권 보호 강화 요구가 확산되고 있다. 개별 창작자들 역시 심각한 영향을 받고 있는데, 일반 사단법인 아츠 워커스 재팬에 따르면 응답자의 90% 이상이 권리 침해를 경험했다고 한다.

출판권과 초상권을 침해하는 개인의 얼굴이나 목소리의 무단 사용 사례가 있다. 일본 출판권 보호기구의 조사에 따르면, 주요 SNS에서 "AI로 다른 사람이 되어 봤다" 또는 "AI로 노래를 불러보게 했다"와 같은 문구가 달린 게시물이 8만 건 이상이며, 총 조회 수는 2억 6천만 회에 달했다. 그 밖에 가수 Drake와 The Weeknd의 목소리를 모방한 AI 생성 곡이 삭제 요청 제기 전까지 1천만 회 이상의 재생 수를 기록한 사건, 배우 Tom Hanks가 등장하는 가짜 광고가 있다. 이 사건들은 해당 개인의 명성과 계약 관계에 심각한 영향을 미쳤다. 일본에서는 배우의 음성 데이터가 허가 없이 학습 데이터로 사용되었고, 관련 협동조합이 '음성권(right of voice)'에 대한 법적 보호를 요구하고 있다. 일반 대중 역시 피해를 입고 있는데, 동의 없이 SNS 게시물이 학습에 사용되거나, 개인이 성적 또는 폭력적 콘텐츠에 등장하는 것처럼 만들어진 사례들이 보고되고 있다. 이는 음란물 생성 및 유통(3.1.4절)에서 논의된 바와 같다. 이와 같은 무단 사용은 개인의 통제를 벗어나 정체성과 인격적 완전성이 희석되는 '인격 상실'이라는 형태의 개인적 피해를 초래하며, 자기 정체성에 대한 확인의 토대를 약화시킨다.

생성형 AI를 활용해 사망한 인물을 재현하는 서비스와 관련된 윤리적 우려가 있다. 이러한 재현은 유가족의 애도 치유에 기여하거나 역사적 인물과 관련된 문화 자산을 보존하는 데 도움이 되나, 인격의 왜곡이나 애도를 장기화시키는 의존성 형성과 같은 위험도 있다. 케임브리지 대학교의 연구는 매우 정교한 AI 재현물이 개인에게 원치 않는 방식으로 영향을 미치고, 심각한 심리적 고통을 초래할 수 있다고 경고한다. 또한 사망한 인물을 이윤 창출을 목적으로 상업적으로 이용하는 것에 대한 우려도 있다. 법적 관점에서 보면, 사망자의 동의를 어떻게 해석할 것인지, 그리고 유가족의 권한을 어떻게 규정할 것인지에 대해 아직 해결되지 않은 쟁점들이 남아 있다. 코넬 대학교의 연구는 생성형 AI가 사망자에 관한 정보를 바탕으로 새로운 콘텐츠를 생성할 수는 있지만, 데이터의 출처와 맥락이 상실될 수 있고, 그 결과 실제 인물과 일치하지 않는 발언이 생성될 위험이 커진다고 지적한다.

AI 생성 결과물과 관련된 권리 문제의 이해관계자로는 AI 개발자, 창작자 및 권리자, 그리고 정부 및 규제 당국이 있다. AI 개발자는 권리와 연관된 학습 데이터의 선정에 대한 책임을 통해 직접적으로 관여한다. 창작자와 권리자는 수익 기회의 상실과 직업적 위상의 약화를 겪을 수 있다. 최종 이용자는 인지하지 못한 채 침해 행위에 기여할 가능성도 있다. 정부 및 규제 당국은 제도적 대응을 추진하고 있다. 예를 들어 일본 정부 문화청은 가이드라인을 발표하였고, 일본신문협회는 성명을 발표하였다. 그러나 국제적 규제 조율은 여전히 과제로 남아 있다.

3.2.2 고용 및 노동시장에 미치는 영향

■ **개요:** 많은 기업들이 업무 효율성을 제고하기 위해 다양한 비즈니스 운영에 생성형 AI를 적용하고 있으며, 이는 고용과 노동시장에 변화를 초래하고 있다. 생성형 AI의 확산은 효율성 향상과 같은 긍정적인 영향을 노동시장에 가져오지만, 인간의 노동을 대체할 경우의 일자리 감소 등 잠재적인 부정적 영향 역시 지적되고 있다. 생성형 AI의 광범위한 도입은 노동시장에 긍정적·부정적 영향을 모두 미칠 것으로 여겨진다. 또한 일자리 감소와 같은 사안은 사회 전반에 중대한 영향을 미칠 수 있으므로, 이에 대한 가능한 대응 방안을 체계적으로 정리한다.

■ **문헌 조사 결과:** 먼저 업무 효율성 향상과 새로운 일자리 창출과 같은 긍정적 영향과, 생성형 AI가 기존 업무를 대체함으로써 발생하는 실업 등 부정적 영향을 모두 보여주는 관련 사례들을 제시한다. 이어서 관련 연구 결과를 소개하고, 마지막으로 이러한 영향으로 인해 영향을 받을 수 있는 이해관계자들을 고찰한다.

첫째, 이러한 영향과 관련된 사례들을 제시한다. NS Solutions Co.에 따르면, 생성형 AI는 번역이나 스프레드시트 수식 작성과 같은 간접 업무를 효율화하여 도입 후 3개월 만에 9,500시간 이상의 근무 시간을 절감하였다. 한편 싱가포르의 은행인 DBS는 AI가 일부 업무를 대체할 것이며, 2028년까지 4,000명의 직원을 감축할 계획이라고 발표하였다. 이러한 사례들은 생성형 AI가 노동시장에 미치고 있는 영향의 다양성을 잘 보여준다.

둘째, 이러한 영향과 관련된 연구 결과를 제시한다. 일본 정부 내각부와 Harvard Business School이 발표한 보고서에 따르면, 생성형 AI가 노동시장에 미치는 영향을 평가할 때 다음의 두 가지 측면을 고려하는 것이 중요하다. 첫 번째 측면은 생성형 AI가 인간의 직무를 완전히 대체하여 인간의 개입 여지가 남지 않는 '대체적' 측면이다. 전통적으로 사람들이 수행해 온 많은 사무 업무는 컴퓨터 성능의 향상으로 이미 노동 집약도가 낮아져 왔다. 생성형 AI의 도입으로 이러한 사무 업무는 더욱 효율화될 수 있으며, 일부는 거의 전면적으로 자동화될 가능성도 있다. 이처럼 과업 수행에 인간의 참여가 더 이상 필요하지 않게 될 경우, AI는 사실상 노동자를 대체하게 된다. 두 번째 측면은 AI가 인간의 업무 부담을 완화하고 생산성을 높이며, 나아가 새로운 일자리 창출을 촉진할 수도 있는 '보완적' 측면이다. MIT 연구진의 실험에 따르면, AI를 활용할 경우 보고서나 이메일 작성과 같은 과업에서 생산성이 향상되는 것으로 나타났으며, 이는 AI가 인간의 과업과 직무를 보완할 수 있음을 시사한다. 즉, AI는 노동자의 업무 중 일부를 담당함으로써 인간과 AI가 협업할 수 있는 환경을 가능하게 한다.

대체적 성격이 강한 직무가 많은 산업에서는 생성형 AI가 효율성 제고를 통한 비용 절감과 같은 상당한 긍정적 효과를 가져올 수 있지만, 동시에 해당 직무 종사자들의 실업과 같은 심각한 부정적 결과를 초래할 수도 있다. 보완성이 높은 직무라 하더라도 일부 과업은 효율화되거나 자동화될 수 있으므로, 고용 규모가 일정 부분 감소할 가능성은 존재한다. 그럼에도 불구하고 생성형 AI가 전혀 새로운 유형의 일자리를 만들어낼 것이다.

생성형 AI의 광범위한 도입으로 인한 노동시장 변화와 관련된 이해관계자로는 AI 개발자, AI 제공자, AI 사용자, 그리고 최종 이용자가 있다. AI 개발자는 많은 기업이 다양한 비즈니스 운영에 AI를 적용할 수 있는 능력이 AI 모델의 성능에 크게 의존한다고 여겨진다는 점에서 관련성을 가진다. AI 제공자는 AI 시스템을 애플리케이션, 제품, 기존 시스템, 업무 프로세스 등에 내재화하기 위해 AI 시스템을 기업의 내부 운영과 통합하는 역할을 수행하므로, 중요한 이해관계자로 간주된다. AI 사용자와 최종 이용자는 정보와 데이터를 다루는 사무 업무가 핵심인 산업에서 중요한 이해관계자이다. 이는 일본 내외에서 수행된 연구들이 산업 및 직종별 노동 보완성 비율을 추정하여 노동의 대체 가능성을 평가해 왔으며, 앞서 언급한 사무 중심 산업의 과업들은 AI가 강점을 보이는 영역으로, 노동 대체 가능성이 높기 때문이다.

■ **인터뷰 조사 결과:** 금융 서비스 분야 기업 G와 IT 서비스 분야 기업 H를 인터뷰하였다.

● **기업 G와 기업 H 인터뷰 결과를 바탕으로 한 고찰:** 인터뷰를 통해, AI가 다양한 사무 업무에 폭넓게 적용되면서 인간의 업무가 효율화되고 있음을 확인하였다. 현재 AI의 역할은 주

로 인간의 노동을 보완하는 데 있다. 효율성 향상뿐만 아니라, AI는 직원의 동기를 높일 수 있는 정서적 가치도 제공할 수 있는 것으로 보인다. 그러나 AI 성능이 지속적으로 향상됨에 따라, 과거에는 사람이 수행하던 과업이 AI로 대체될 가능성도 존재한다. 이에 따라 직원들의 AI 리터러시를 제고하고, AI 활용을 전제로 업무 프로세스를 재설계하는 노력이 필요하다. 또한 AI 도입을 더욱 촉진하기 위해서는 환각과 같은 AI 고유의 위험을 적절히 관리해야 한다.

3.2.3 생성 콘텐츠 확산의 영향

■ **개요:** 생성형 AI는 저비용과 고속으로 대량의 메시지와 콘텐츠를 생성한다. ChatGPT와 같은 도구는 단일 프롬프트만으로도 인간이 제작한 것에 필적하는 텍스트와 이미지를 생성할 수 있으며, 패러프레이징 기능을 통해 동일한 내용을 무수한 변형으로 표현한다. 이러한 특성은 설문조사, 리뷰, 소셜미디어 게시물 등 다양한 맥락에서 품질보다 양을 우선하는 게시물이 기존의 스팸 탐지 방식을 우회하도록 하였다. 그 결과, 대량의 저품질 정보가 범람하고 있다. 일부 사례에서는 순위 결정 알고리즘이 이러한 콘텐츠를 상위에 노출시켜, 고품질 자료의 가시성이 저하되는 것으로 보고되었다. 이에 따라 플랫폼들은 기존의 패턴 매칭 기반 탐지 방식에서 벗어나, 생성형 AI를 활용한 새로운 스팸 탐지 방식으로 전환할 수밖에 없는 상황에 놓이고 있다. 생성 콘텐츠의 확산은 사회적으로 큰 영향을 미칠 수 있기 때문에, 본 프로젝트에서는 이 주제를 연구 과제로 규정한다. 앞서 설명한 부정적 영향뿐만 아니라, 생성 콘텐츠의 확산은 가짜 리뷰와 같은 기존 문제에 대응하기 위한 규제 정비를 가속화하는 등 긍정적인 효과도 가져오고 있다. 이하에서는 생성 콘텐츠 확산과 관련된 부정/긍정적 측면을 정리한다.

■ **문헌 조사 결과:** 생성 콘텐츠의 확산은 플랫폼의 예상을 뛰어넘는 규모로 발생하여, 설문조사와 리뷰의 신뢰성 저하, 소셜미디어 생태계의 오염, 검색 서비스 품질의 악화와 같은 부정적 측면뿐만 아니라, 규제 정비가 가속화되는 등의 긍정적 측면도 함께 관찰되고 있다.

부정적 영향과 관련한 사례를 살펴보자. 첫째, 설문조사와 리뷰의 신뢰성이 악화되고 있다. 스탠퍼드대학교의 Janet Xu 교수에 따르면, 온라인 설문조사 참여자의 약 3분의 1이 ChatGPT와 같은 AI 도구를 사용해 응답을 작성한 것으로 나타났다. 그 결과 데이터의 동질화가 발생하여, “응답에 오타가 더 적고 … 지나치게 친절해 보였다”거나 “LLM은 인종, 정치, 기타 민감한 주제에 대해 더 중립적이고 추상적인 언어를 일관되게 사용해, 보다 거리감을 두고 접근하는 경향을 보였다”와 같은 특징이 나타났다. 이러한 데이터는 연구와 정책 수립의 기초 자료로서의 신뢰성을 약화시킬 수 있다. 둘째, 소셜미디어 생태계가 오염되고 있다. 하버드 케네디 스쿨과 스탠퍼드대학교의 공동 연구는 생성형 AI로 제작된 이미지를 활용해 페이스북에서 급속히 확산된 스팸 페이지들을 보고했으며, 단일 게시물이 수백만 건의 반응을 얻은 사례도 확인되었다. 또한 언론 보도에 따르면, 개발도상국의 창작자들이 높은 광고 수익을 얻기 위해 생성형 AI를 활용해 미국 시장을 겨냥한 콘텐츠를 대량 생산하고 있는 실태도 드러났다. 최종 이용자는 이러한 콘텐츠가 AI 생성물임을 인지하지 못한 채 소비하게 되며, 그 결과 플랫폼 전반의 정보 환경이 오염된다. 셋째, 검색 서비스의 품질이 저하되고 있다. 구글 검색에서 저품질 콘텐츠의 비중이 증가하고, 상위 검색 결과가 AI 생성물로 지배되고 있는 상황이다. 이러한 지배 현상에 대응하기 위해 구글은 여러 차례 ‘spam update’ 조치를 시행하고 있다. 검색의 신뢰성이 저하될 경우, 그 영향은 인터넷 전체의 정보 이용 인프라로까지 확산된다.

긍정적인 측면에서 보면, 생성형 AI의 확산은 규제 강화를 촉진하는 계기로 작용하였다. 생성형 AI 확산 이전에도 가짜 리뷰는 존재했지만, 그 규모는 제한적이었고 엄격한 규제도 거의

없었다. 그러나 AI가 생성한 리뷰 게시물이 대규모로 증가하면서 소비자 피해와 시장 왜곡이 확대되자, 미국 FTC는 2024년에 AI가 생성한 것을 포함한 가짜 리뷰가 규제 대상에 해당하며, 위반 시 민사상 제재금이 부과될 것임을 명확히 하였다. 이러한 조치는 생성형 AI가 초래하는 위험에 법 제도가 대응해 나가기 위한 중요한 출발점으로 평가할 수 있다.

생성 콘텐츠 확산의 영향과 관련된 이해관계자로는 AI 개발자, AI 제공자, 그리고 최종 이용자가 있다. AI 개발자는 악의적 행위자가 스팸을 생성할 수 있도록 하는 도구를 제공하는 한편, 플랫폼의 스팸 탐지를 지원하는 엔지니어 역할도 수행할 수 있어, 공격과 방어 양 측면에 모두 관여한다. AI 제공자 중에서 검색 엔진, 소셜 네트워크 서비스, 리뷰 사이트, 순위 결정 및 설문 플랫폼 등 자사 서비스에 AI 기능을 내재화한 플랫폼 운영자가, 생성형 AI로 만들어진 스팸 게시물에 직접 노출되는 주요 이해관계자가 될 가능성이 크다. 그 결과 사용자 경험이 저하되고, 광고 수익에 부정적 영향을 미쳐 플랫폼 운영의 지속 가능성을 위협할 수 있다. 최종 이용자 중에서는 스팸을 게시하는 이들이 보상이나 이익을 얻기 위해 생성형 AI를 악용할 수 있으며, 익명성과 즉시성이 이러한 행위를 뒷받침한다. 이러한 동기는 개인에서 조직에 이르기까지 다양하며, 도구의 사용이 쉬울수록 오·남용에 대한 유인이 더욱 강해진다.

3.2.4 경제적 불평등에 미치는 영향

■ **개요:** 생성형 AI가 사회·경제 활동 전반을 변화시킬 수 있는 범용 기술로서 지니는 역량에 대한 관심이 집중되고 있다. 기존 AI가 주로 분석과 의사결정을 담당해 왔던 것과 달리, 생성형 AI는 '창작'의 영역까지 확장되었다. 그 결과, 한때 인간만의 영역으로 여겨졌던 창의적 활동까지 자동화 되어 생산성의 획기적인 향상이 기대되지만, 개인의 소득이나 기업의 시장 가치 등과 관련하여 이후 논의될 여러 우려도 제기되고 있다. 경제적 불평등의 관점에서 볼 때, 생성형 AI의 혜택이 사회 전반에 공평하게 분배되지 않고, 특정 개인·기업·국가에 부가 집중되어 기존의 격차가 확대되거나 새로운 불평등이 형성될 수 있다는 우려가 제기되고 있다.

■ **문헌 조사 결과:** 생성형 AI의 확산은 경제적 불평등에 여러 측면에서 영향을 미칠 수 있다. 본 절에서는 개인(직업), 기업, 국가 차원에서의 영향을 다룬다.

개인(직업) 차원에서는 3.2.2절에서 언급한 바와 같이, 생성형 AI는 사무 업무, 콜센터 업무, 텍스트 또는 코드 생성과 같은 반복적 과업에 대해 높은 대체 가능성을 지닌 것으로 평가된다. 본 절에서는 이러한 맥락을 바탕으로 경제적 불평등에 초점을 맞추어 논의를 전개한다. IMF의 연구에 따르면, 노동소득에 대한 영향은 생성형 AI와의 보완성에 따라 달라진다. 생성형 AI를 보완하는 기술과 역량을 가진 노동자는 소득이 증가할 가능성이 있는 반면, 생성형 AI에 의해 쉽게 대체될 수 있는 직종은 소득 증가가 제한될 수 있다. 그 결과 전체적인 생산성이 향상되더라도 그 성과는 고르게 분배되지 않아 노동소득 격차가 확대될 수 있다. 동시에 생성형 AI에 투자하거나 관련 자산을 보유한 이들에게 이윤이 집중됨에 따라, 자본소득과 자산 격차 역시 확대될 가능성이 있다. 반대로 MIT의 연구는, 생성형 AI가 고소득층에게 요구되던 관리 업무나 고급 개발 업무를 대체할 경우 이러한 노동의 상대적 가치가 하락하여, 오히려 격차가 축소될 가능성도 있다고 지적하고 있다.

AI 개발에 필요한 막대한 연산 자원, 데이터, 인재를 확보할 수 있는 대규모 기술 기업이 소수에 불과하다. 이러한 기업들이 제공하는 플랫폼에 대한 의존도가 높아질 경우, 기업 간 격차는 더욱 확대될 수 있다. 생성형 AI의 업무 활용 측면에서도 대기업과 중소기업 간 도입 및

활용 수준의 차이가 생산성과 경쟁력의 격차로 이어질 가능성이 있다. 일본의 경우, 대기업의 절반 이상이 생성형 AI 활용 방침을 마련하고 있는 반면, 중소기업에서는 그 비율이 약 30%에 불과한 것으로 나타났으며, 이러한 차이는 향후 생산성과 경쟁력 격차로 전이될 수 있다. 미국의 경제 연구기관 보고서에 따르면, 생성형 AI의 진전은 기술 중심 기업과 전통 산업 기업 간의 주가 격차를 가속화하고 있으며, 이는 시장 가치 평가의 격차 확대를 시사한다. 이러한 추세는 재정적 여력과 데이터 기반을 갖춘 기업에 더욱 유리하게 작용한다.

국가 차원에서는 디지털 무역수지가 주요 과제로 지적된다. 디지털 무역수지란 광고비, 소프트웨어 라이선스 비용, 클라우드 서비스 이용료, 로열티 등 디지털 관련 거래에서의 상태를 의미한다. 일본의 경우 2024년에 이미 약 6조 엔 규모의 적자를 기록했으며, AI 관련 서비스의 상당수가 해외 기업에 의해 제공되고 있는 점을 고려하면, 2035년에는 그 적자 규모가 약 28조 엔에 이를 것으로 전망되고 있다. 즉, 생성형 AI의 확산과 함께 해외에서 제공되는 서비스에 대한 의존도가 높아질 경우, 소득의 해외 유출과 국제 경쟁력 저하를 초래한다.

경제적 불평등에 미치는 영향과 관련된 이해관계자로는 AI 개발자와 AI 제공자, AI 사용자/최종 이용자, 교육·훈련 기관, 그리고 정부 및 규제 당국이 있다. AI 개발자는 최종 이용자와 기업이 지불하는 이용 요금의 상당 부분이 수익으로 귀속된다는 점에서 관련된다. AI 제공자는 생성형 AI를 애플리케이션과 업무 프로세스에 내재화하여 경제적 가치를 창출하는 주체로서, 특히 내부 도구를 제공하는 플랫폼 운영자와 통합 시스템을 제공하는 시스템 통합 사업자가 이에 해당한다. AI 사용자/최종 이용자는 정보와 데이터를 집중적으로 다루는 사무 중심 산업—예를 들어 정보통신, 금융·보험, 교육 및 학습 지원 분야—에서 특히 큰 영향을 받는다. 정부 및 규제 당국은 생성형 AI로 인한 경제적 불평등 문제에 대응하고, 공공의 AI 리터러시를 제고하며, 개발 및 활용 정책을 수립하는 등 다각적인 정책 대응을 추진할 것으로 기대된다.

3.2.5 기밀 정보의 학습 및 유출에 대한 우려

■ **개요:** 생성형 AI에 의한 기밀 정보의 학습 및 유출 위험이 지적되고 있다. 대규모 학습 과정에서 생성형 AI가 기밀 정보나 개인정보를 포함한 데이터를 흡수할 가능성이 있으며, 최종 이용자가 입력한 정보가 재사용될 수도 있다. 그 결과 생성 과정에서 기밀 정보가 외부로 출력되어 개인정보 침해, 기업 경쟁력 약화, 법적 책임과 같은 위험을 초래할 수 있다. 기밀 정보의 학습 및 유출에 대한 우려는 최종 이용자에게만 국한되지 않으며, 기업·정부·사회 전반에 대한 신뢰를 훼손할 수 있다. 기밀 정보의 관리와 보호는 사회 전체 차원의 대응이 필요하다.

■ **문헌 조사 결과:** 기밀 정보의 학습 및 유출로 인한 영향은 다면적이다. 이 문제를 기업 경쟁력/신뢰의 약화, 개인 프라이버시 침해, 법규·규정 준수 위험의 증가로 나누어 정리한다.

첫째, 기업 경쟁력과 신뢰의 약화와 관련하여, 2023년 삼성전자 직원이 기밀 내부 소스 코드를 ChatGPT에 입력한 사례가 보고되었다. 입력된 정보는 외부 서버에 저장되어 삭제가 어렵고, 다른 이용자에게 공개될 가능성도 있었기 때문에, 삼성전자는 생성형 AI 사용을 금지하는 새로운 정책을 수립하였다. 또한 ChatGPT가 기본적으로 대화 기록을 저장하고 이를 학습에 활용한다는 점 역시 위험 요인으로 지적되었다. 이 사례는 편의성을 추구하는 과정에서 기업의 지식재산과 전략이 외부로 유출될 수 있는 위험성을 잘 보여준다. 둘째, 개인정보 침해와 관련하여, 2023년 OpenAI에서 시스템 사고가 발생해 약 9시간 동안 일부 이용자의 이메일 주소, 청구지 주소, 신용카드 번호의 마지막 네 자리와 유효기간이 다른 이용자에게 노출되는

문제가 발생했다. 해당 사고의 원인은 오픈소스 라이브러리의 결함으로 확인되었고, 영향을 받은 이용자들에게는 통지가 이루어졌으나, 이 사건으로 인해 이용자들 사이의 신뢰는 크게 흔들렸다. 널리 사용되는 생성형 AI를 통해 개인정보가 유출될 경우, 피해가 즉각적으로 확산되어 사회적 불안을 증폭시킬 수 있다. 셋째, 법규 및 규정 준수 위험과 관련하여, 2023년 미국에서 OpenAI와 Microsoft가 인터넷에 공개된 개인정보를 불법적으로 수집하고 이를 학습에 활용했다는 혐의로 소송이 제기되었다. 2024년에는 캘리포니아 연방지방법원이 소장 기재의 하자를 이유로 해당 소송을 각하했으나, 이러한 소송이 제기되었다는 사실 자체가 AI 개발자에게 중대한 법적 위험이 존재함을 시사한다. EU의 일반개인정보보호규정과 같이 엄격한 개인정보 보호 체계 하에서는 위반 시 막대한 과징금 부과나 사업 중단으로 이어질 수 있다.

이러한 사례들은 기밀 정보의 학습 및 유출 문제가 기업 경영, 개인의 일상생활, 나아가 국가 안보에까지 영향을 미칠 수 있는 사회적 문제임을 보여준다. 일본에서는 일본디지털경제·커뮤니티진흥기구와 ITR이 일본 기업을 대상으로 실시한 조사에서, 생성형 AI 활용과 관련한 가장 큰 우려 사항으로 「사내 기밀 정보를 학습 데이터로 사용하는 데서 발생하는 정보 유출」이 꼽혔다. 또한 OWASP Top 10 for LLM Applications 2025에서는 주요 위험 요소 중 하나로 「민감 정보 노출」이 포함되어 있다. 이는 기밀 정보의 학습 및 유출에 따른 위험이 사회 전반에 걸쳐 널리 인식되고 있으며, 이에 대한 대응이 시급하다는 점을 뒷받침한다.

기밀 정보의 학습 및 유출에 대한 우려와 관련된 이해관계자로는 AI 개발자, AI 제공자, 그리고 AI 사용자/최종 이용자가 있다. AI 개발자는 모델 설계와 학습 단계에서 입력 데이터가 학습에 사용되는지 여부와 데이터가 어떻게 저장·활용되는지를 결정한다. AI 제공자는 이용 약관, 출력 제어, 접근 통제, 콘텐츠 모더레이션 등을 통해 정보 유출을 방지하는 데 직접적으로 관여한다. AI 사용자/최종 이용자 역시 핵심적인 이해관계자이다. 특히 국가 안보와 핵심 인프라, 금융, 의료 등과 같이 고도의 민감 정보를 다루는 산업은 중요한 이해관계자이다.

3.3 정보 공간에 대한 영향

3.3.1 허위 정보 및 조작 정보의 생성과 확산

■ **개요:** 생성형 AI의 특성이 악용되어 가짜 기사나 허위 뉴스 영상이 짧은 시간 안에 제작될 수 있게 되었다. 또한 소셜 미디어의 확산으로 인해 이러한 조작 정보는 손쉽게 전파될 수 있으며, 그 결과 사회적 혼란을 초래할 가능성이 있다. 더 나아가 알고리즘 기반 추천과 커뮤니티의 동질화가 결합되면서, 서로 상충되는 정보를 접할 기회가 줄어들고 있다. 이는 개인이 자신의 선입견이나 의견을 뒷받침하는 정보만을 수집하게 만들어, 이른바 '에코 챔버 현상'을 강화한다. 에코 챔버 현상이란 유사한 가치관을 가진 최종 이용자들이 서로의 공감을 강화함으로써 특정 의견이나 이데올로기가 과도하게 증폭되고 그 영향력이 커지는 상태이다. 아울러 생성된 콘텐츠가 요약본, 회의록, 보고서 등 직장에서 공유되는 '중요 문서'에까지 침투할 가능성도 있으며, 이로 인해 의도하지 않은 허위정보가 의사결정에 영향을 미칠 위험이 증가하고 있다. 생성형 AI를 통한 허위 정보 및 조작 정보의 생성과 확산은 생성형 AI의 기술적 특성과 정보사회 구조 간의 상호작용에 의해 증폭되는 사회·기술적 문제이다.

■ **문헌 조사 결과:** 「기사와 영상 등을 통한 허위 정보 및 조작 정보의 확산」, 「생성형 AI에 의한 에코 챔버 현상의 촉진」, 「중요 문서로의 허위정보 및 조작정보 유입」 관점에서 다룬다.

첫째, 생성형 AI가 오·남용되어 만들어진 허위정보/조작정보의 확산은 선거, 시장, 재난 대응

과 같은 사회의 근본적인 영역에 심각한 영향을 미치고 있다. Nikkei의 보도에 따르면, 2024년 이후 대만과 일본의 선거 과정에서 후보자와 총리를 사칭한 가짜 영상이 확산되는 사례가 확인되었다. 또한 2023년에는 중국기업 iFlytek의 주가가 조작정보로 인해 일시적으로 9% 하락하며 시장에 혼란을 초래한 바 있다. 재난 상황에서도 예를 들어, 시즈오카현에서 발생한 태풍 제15호 당시 AI가 생성한 가짜 이미지가 유통되어 대피 및 구호 활동을 방해한 사례가 보고되었다. 더 나아가 소셜미디어 플랫폼은 클릭률과 참여도가 높은 게시물을 우선적으로 노출하는 경향이 있어, 사회적 파급력이 큰 가짜 기사가 구조적으로 사실 보도보다 더 확산될 가능성이 있다. MIT의 연구에 따르면 거짓 뉴스는 진실한 뉴스보다 약 70% 더 빠르게 확산되며, 1,500명에게 도달하는 데 걸리는 시간도 진실한 정보의 6분의 1에 불과한 것으로 나타나, 소셜미디어의 구조적 특성이 허위정보 및 조작정보 확산에 기여하고 있음을 보여준다.

둘째, 생성형 AI에 의한 에코 챔버 현상 촉진과 관련하여, 생성형 AI의 대화형 특성과 고도화된 개인화 기능은 정보 편향을 심화시킬 위험을 내포하고 있다. 미국에서는 선택적 정보 노출과 의견 양극화가 빠르게 진행될 수 있으며, 이를 교정하려는 조치가 대부분 효과를 거두지 못하였다. 일본 총무성의 「2024년 정보통신 백서」와 디지털청의 「공공행정의 발전과 혁신을 위한 일본 정부의 생성형 AI 조달 및 활용 가이드라인」에서도 생성형 AI가 에코 챔버 효과를 심화시킬 위험이 있음을 경고하고 있다. 따라서 개인 차원에서는 의견 경직화와 인지 편향이 나타날 수 있으며, 사회 차원에서는 양극화 심화와 공적 담론의 축소가 발생할 수 있다. 또한 정부 차원에서는 행정과 정책 결정의 편향화로 이어져 다면적인 영향을 초래할 수 있다.

셋째, 중요 문서로의 허위정보 및 조작정보 유입은 사회 신뢰의 기반에 심각한 영향을 미칠 수 있다. 학계에서는 생성형 AI가 만들어낸 가짜 논문 사례가 유통된 바 있으며, 사법 분야에서는 AI가 생성한 존재하지 않는 판례를 잘못 인용한 사례도 보고되었다. 기업 분야에서는 재무 문서나 계약서에서 발생한 생성형 AI의 오류가 경영 판단이나 주가에 직접적인 영향을 미칠 수 있다. 특히 의료, 공공 안전, 환경 정책 등 생명과 일상생활과 직접적으로 관련된 영역에서는 허위 정보의 유입이 사회적 공황이나 잘못된 행동을 초래할 수 있다.

생성형 AI를 통한 허위 정보 및 조작 정보의 생성/확산과 관련된 이해관계자로는 AI 개발자, AI 제공자 및 소셜미디어 플랫폼, 학술 기관 및 언론사, 정부, 그리고 최종 이용자가 있다. AI 개발자는 학습 데이터와 출력에 허위정보가 반영될 경우, 잘못된 정보를 재생산할 위험에 직면한다. AI 제공자와 소셜미디어 플랫폼은 허위정보의 유통 과정에 관여한다. 학술 기관과 언론사는 가짜 논문이나 허위 보도의 게재를 통해 연구 결과와 뉴스의 전반적인 신뢰성을 훼손당할 수 있다. 정부는 허위정보가 정책 결정과 의사결정 과정에 침투할 경우, 공공 생활과 사회 시스템에 피해를 입을 위험이 있다. 또한 최종 이용자는 허위정보의 피해자일 뿐만 아니라, 소셜미디어와 같은 플랫폼을 통해 무의식적으로 정보를 확산시키는 역할을 하기도 한다.

■ **인터뷰 조사 결과:** 비영리 팩트체크 조직인 회사 I와 언론사인 회사 J를 대상으로 하였다.
● **회사 I와 회사 J 인터뷰 결과를 바탕으로 한 고찰:** 인터뷰 대상 조직들은 AI에 의해 생성된 허위정보와 조작정보가 증가하고 있음을 인식하고 있다. 그러나 AI 생성 허위정보에 대한 통계 자료를 확보하기 어렵다는 점도 지적하며, 향후 피해의 실태를 조사하는 것이 중요하다고 언급하였다. 또한, 단순히 “인간의 판단”에만 의존하거나 검증을 일반 대중에 맡기는 것에는 한계가 있다. 따라서 인간의 판단과 기술적 지원을 결합한 하이브리드 접근이 필요하다. 동시에 모든 이해관계자는 미디어 리터러시 교육, 법적 체계, 산업 표준, 신뢰의 기술적 증명 구축

등 대응책을 강화하고, 상호 협력을 통해 함께 대응할 필요가 있다. 이러한 노력은 조작정보의 확산을 방지하고 보다 안전한 정보 사회를 촉진하는 것과 직접적으로 연결된다.

3.3.2 다양성에 대한 영향

■ **주제 개요:** AI가 생성한 결과물이 충분히 다양성을 반영하는지에 대한 우려가 있다. 학습 데이터가 특정 속성이나 문화에 편향되어 있는 경우, 유사한 편향이 결과물에도 반영될 수 있어 사회적 다양성을 저해할 가능성이 있다. 예를 들어, 인종, 성별, 연령, 장애 등 속성과 관련된 편향이나, 지역, 언어, 종교 등 문화적 다양성과 관련된 편향이 관찰된 바 있다. 그 결과, AI가 생성한 결과물에서 사회적으로 소외된 집단이 보이지 않게 되거나, 기존의 고정관념에 제한된 형태로 표현될 수 있다. 동시에, 적절하게 설계되고 적용된다면 생성형 AI는 의사소통과 업무 수행의 장벽을 낮추고, 다양한 인재가 역량을 발휘할 기회를 확대하여 긍정적인 효과를 낼 수 있다. 이에 생성형 AI가 다양성에 미치는 부정적 측면/긍정적 측면을 정리한다.

■ **문헌 연구 결과:** UNESCO의 발간물을 바탕으로, 다양성을 사람, 문화, 언어의 다양한 차이를 인식하고 존중하는 것으로 정의하고, 속성/문화의 다양성에 중점을 둔다.

속성의 다양성: 독립 사회과학 연구자 Sadeghiani는 이미지 생성 AI를 활용한 실증 연구를 수행하고, 직업과 관련된 444개의 이미지를 분석하였다. 연구 결과, 흑인, 여성, 고령자, 장애인 등 특정 속성이 현저히 과소 대표되는 것으로 나타났다. 여성은 과학 및 기술 분야에서 거의 묘사되지 않았으며, 눈에 띄는 장애를 가진 사람은 단 한 번도 묘사되지 않았다. 중장년층과 노년층은 고정적으로 제한된 역할로만 표현되었다. 이러한 결과는 생성형 AI가 학습 데이터에 존재하는 편향을 재생산하고, 사회 속 속성 다양성을 저하시킬 위험이 있음을 시사한다. 반면, 생성형 AI가 다양성을 지원하는 긍정적 측면도 보고된 바 있다. Ernst & Young의 국제 조사에 따르면, 장애인 또는 신경다양성(neurodiversity)을 가진 직원의 85%가 “생성형 AI 도구가 직장을 보다 포용적으로 만들고 있다”라고 응답했다. 예를 들어, 실시간 전사 및 자동 요약 기능은 청각 장애인이나 발달 장애인을 지원하고, 글쓰기 보조 기능은 생각을 구조화하는 데 어려움을 겪는 사람들의 부담을 줄여준다. 적절히 활용될 경우, 생성형 AI는 다양한 인재의 업무 수행을 지원하고 포용적인 직장 환경 조성에 기여할 수 있다.

문화의 다양성: 서구 중심적 가치관에 의해 형성된 AI 모델의 영향에 대한 우려가 제기되고 있다. 전 세계적으로 사용되는 많은 모델이 유럽과 미국에서 개발되기 때문에, 비서구 문화와 언어는 과소 대표될 수 있다. 스탠포드 대학의 Abid 연구팀은 대형 언어 모델이 무슬림을 테러리스트로 묘사하는 경향이 있음을 지적하며, 비서구 문화에 대한 표현적 피해를 보여주었다. 또한 인도 참가자의 경우, AI의 제안에 의존하면 서구식 글쓰기 방식을 채택하게 되고 문화적 균질화가 촉진될 수 있다고 경고했다.

다양성에 대한 영향과 관련된 이해관계자는 AI 개발자, 교육 기관 및 창작 산업, 정부 및 국제기구 등이 있다. 특정 속성(여성, 흑인, 장애인 등)의 과소대표가 이미 관찰되고 있기 때문에, AI 개발자는 모델 설계에서 편향을 완화할 책임을 져야 한다. 교육 기관과 창작 산업은 다양성이 결여된 콘텐츠가 광범위하게 확산될 경우 학습자와 사회에 편견을 재생산할 위험에 직면하며, 부적절한 표현은 장애인 및 소수자 커뮤니티를 대표하는 조직의 권한 강화 활동을 저해할 수도 있다. 거버넌스 주체 또한 필수적인 역할을 수행한다. 정부는 다양한 이해관계자를 조율하고 학제 간 이니셔티브를 촉진한다. 국제기구인 UNESCO는 AI 결과물에서 속성 관

련 다양성이 부족하다는 조사 결과를 발표했으며, AI 윤리에 관한 권고안을 통해 AI 도구 설계에서 성평등을 확보하기 위한 구체적 행동을 촉구하였다. 생성형 AI의 확산이 가속화됨에 따라, 국제 표준과 가이드라인을 통한 프레임워크 구축은 여전히 중요할 것이다.

3.4 환경에 대한 영향

3.4.1 환경에 대한 영향

■ **주제 개요:** 생성형 AI는 기존 소프트웨어에 비해 더 많은 에너지를 소비하므로, 환경에 부정적인 영향을 미칠 수 있다. 미국 전력연구소의 연구에 따르면, 일반적인 구글 검색은 요청당 평균 0.3 Wh의 전력을 소비하는 반면, ChatGPT 요청은 평균 2.9 Wh가 소요된다. 또한 스탠포드 대학의 보고서에 따르면, GPT-3를 학습시키는 데 약 1,300 MWh의 전력이 필요했으며, 더 발전된 GPT-4 학습에는 이보다 50배 많은 전력이 필요했던 것으로 추정된다. 한편, 생성형 AI가 에너지 효율을 개선하고 재생에너지의 최적 활용에 기여하는 사례들도 보고되고 있다. 따라서 생성형 AI가 환경에 미치는 영향을 부정적/긍정적 측면을 정리한다.

■ **문헌 연구 결과:** 생성형 AI가 환경에 미치는 영향에는 CO_2 배출 증가와 전력망 부담 가중과 같은 부정적 측면과, 재생에너지 활용 촉진 및 효율성 향상과 같은 긍정적 측면이 있다.

우선, 부정적 측면의 두 가지 사례를 살펴본다. 첫째는 주요 기술 기업들의 전력 소비와 CO_2 배출 증가이다. 구글은 CO_2 배출을 감축하겠다는 목표를 설정했지만, 생성형 AI 사용 확대에 따라 2019년부터 2023년 사이에 오히려 48% 증가한 것으로 보고되었다. 마찬가지로 Microsoft에서도 데이터센터 확장으로 인해 2020년 이후 전력 사용에 따른 CO_2 배출이 약 25% 증가한 것으로 전해졌다. 이로 인해 CO_2 배출을 수반하는 전력 소비 감축 목표를 설정하고 추진해 왔음에도 불구하고, 해당 목표 달성이 어려워지고 있다. 둘째는 특정 지역에서의 전력망 부담 문제이다. 아일랜드는 지리적 조건으로 인해 유럽 내 데이터센터 허브가 되었으며, 구글과 같은 기업들이 대규모 시설을 구축해 왔다. IEA의 보고서에 따르면, 2026년에는 아일랜드 전체 전력 수요의 32%가 데이터센터에서 발생할 것으로 예측되며, 아일랜드 전력규제위원회는 신규 연결 제한 등 규제를 강화하는 조치를 취하고 있다. 전력 수요의 급격한 증가는 지역 전력 공급의 안정성을 저해하고, 주민들의 삶과 산업 활동에까지 영향을 미친다.

반면, AI가 환경적 지속가능성에 기여함으로써 긍정적인 측면을 가질 수 있다는 지적도 있다. MIT Technology Review에 따르면, 2024년에 공개된 구글 기상 예측 AI 'GenCast'는 풍속과 풍향을 정확하게 예측하고 풍력 터빈의 운영을 최적화함으로써 발전 효율을 향상시켰다. 또한 UPDATER Inc.와 도쿄대학교의 공동 연구에서는 AI 예측 모델을 활용해 발전소의 향후 24시간 전력 생산량을 예측하여 전력 시장에서 효율적인 거래가 가능해진다고 하였다. 이러한 사례들은 AI가 재생에너지의 안정적인 공급을 보장하는 데 기여할 수 있음을 보여준다.

또한 기업들의 재생에너지 조달이 가속화되고 있다. Microsoft는 생성형 AI 전력 수요 증가에 대응하기 위해 Brookfield Renewable Partners와 기존 계약보다 약 8배 규모의 전력구매계약을 체결하였다. 이는 생성형 AI 사용 확대가 재생에너지 시장을 활성화함을 시사한다.

생성형 AI가 환경에 미치는 영향과 관련된 이해관계자로는 AI 개발자, AI 제공자, AI 이용자 및 최종 이용자, 데이터센터 운영자, 전력 회사 등이 있다. 생성형 AI를 운영하기 위해 막대한 전력을 소비하는 AI 개발자는 에너지 효율적인 AI 모델 개발 방법과 모델 최적화 기술을 구

축할 책임이 있다. 또한 생성형 AI 확산으로 데이터센터의 전력 수요가 급증함에 따라, 메타와 구글 등 주요 기술 기업들이 원자력 발전에 대규모 투자를 하고 있다는 보고도 있다. AI 제공자는 생성형 AI 사용 규모와 시스템 설계가 전력 소비에 큰 영향을 미칠 것으로 예상된다. AI 이용자와 최종 이용자의 AI 사용이 증가할수록 전력 소비도 늘어나지만, 개인이 자신의 사용이 환경에 미치는 영향을 파악하기는 어렵기 때문에, 주로 사용량을 통해 간접적으로 관여하는 것으로 볼 수 있다. 데이터센터 운영자는 인프라 제공자로서 전력 소비 및 냉각 효율과 직접적으로 연결되어 있으며, 입지, 건물 구조, 설비 설계에 따라 지역 환경 부담이 크게 달라질 수 있기 때문에 환경 영향과 잠재적으로 밀접한 관련이 있다. 전력 회사는 AI 관련 수요 증가로 새로운 시장 기회를 얻고 있으며, AI 개발자 및 데이터센터 운영자와의 협력을 통해 재생에너지 도입을 촉진하고 공급 시스템을 강화하는 역할을 수행할 것으로 기대된다.

4 향후 논의를 향하여: 현재의 쟁점과 대응방안

4.1 AI 안전성의 사회·기술적 영향과 관련된 현재의 쟁점

생성형 AI의 급속한 확산은 범죄적 오남용과 법적 문제에서부터 경제적 불평등과 같은 보다 광범위한 사회·구조적 변화에 이르기까지 다양한 문제를 드러낼 수 있다. 본 절에서는 이러한 쟁점을 기술적 문제, 사회적 문제, 제도적 문제로 구분하여 논의한다.

■ **기술적 쟁점:** 이와 관련하여 두 가지 사항, 즉 생성형 AI의 학습 데이터 품질과 부적절한 출력이 미치는 다양한 영향을 다룬다. 먼저 학습 데이터의 품질에 대해 살펴본다. 일부 생성형 AI 모델은 인터넷 문서 등 방대한 자료를 무작위로 수집한 코퍼스를 기반으로 학습된다. 모델의 출력은 학습 데이터에 크게 의존하기 때문에, 데이터의 품질은 생성 결과의 품질과 직접적으로 연결된다. 3.1.2절에서 지적된 바와 같이, 최종 이용자가 특정 인물의 이미지를 제공하거나 모델을 해당 인물로 학습시키지 않았음에도 불구하고, 텍스트 프롬프트만으로 의도치 않게 그 인물을 연상시키는 외설적인 이미지를 AI가 생성한 사례가 보고되었다. 이는 AI 개발자가 모델 학습을 위해 무작위로 데이터를 사용함으로써, 당사자의 동의 없이 특정 인물의 이미지가 학습 데이터에 포함되었을 가능성을 시사한다. 또한 아동을 대상으로 한 외설물은 전 세계적으로 아동 성적 학대와 직결되는 심각한 문제이다. 마찬가지로 3.1.4절에서는 이러한 자료를 전혀 포함하지 않는 데이터만을 사용해 모델을 학습시켜야 한다는 요구가 제기되었다. 3.1.2절에서 보듯이, 멀티모달 데이터셋의 규모가 확대될수록 생성형 AI의 출력이 인종이나 민족과 같은 사회적 속성에 대해 차별을 드러낼 가능성이 높아진다고 경고하고 있다. 「비즈니스를 위한 AI 가이드라인(Version 1.1)」 역시 학습 데이터를 편향의 요인으로 지적하며, AI 개발자가 데이터의 품질을 관리하고 보장하기 위한 실질적인 조치를 취할 것을 요구한다.

부적절한 출력이 미치는 다양한 영향을 살펴보자. 생성형 AI는 활용 범위가 매우 넓어 자연어 생성/번역/프로그래밍/이미지 생성 등 다양한 작업을 수행한다. 이는 편의성을 높이고 활용 가능성을 확장시키는 한편, 부적절한 출력이 다양한 영향을 미칠 가능성도 함께 증대시킨다. 예) 3.1.3절과 3.3.1절에서 지적된 바와 같이, 악의적인 활용을 통해 피싱 이메일, 위조 기사, 악성 딥페이크와 같은 부적절한 출력이 생성될 수 있다는 우려가 있다. 3.1.1절과 3.2.5절에서는, 최종 이용자가 부적절한 안내를 받아 정신적/신체적으로 부정적 결과를 겪은 사례와, AI가 생성한 출력물을 통해 기밀 정보가 노출된 사례가 보고되었다. 부적절한 출력이 매우 다양한 영향을 미칠 수 있으며, 출력 통제가 중요한 쟁점이다. 따라서 혁신을 촉진하는 동시에, AI 개발자는 오남용, 잘못된 안내, 기밀 정보 유출을 방지하기 위해 모델 출력을 관리해야 한다.

■ **사회적 쟁점:** 두 가지 구체적인 문제, 즉 불평등의 확대와 개인의 리터러시 부족을 다룬다. 먼저 불평등의 확대에 대해 논의한다. 3.2.4절에 따르면, 생성형 AI가 사회 전반으로 확산되고 있음에도 불구하고 이를 적극적으로 활용하는 주체(개인, 기업, 국가))와 그렇지 못한 주체가 존재한다. 생성형 AI의 활용이 사회 전반에서 더욱 가속화될 것으로 예상됨에 따라, AI 도입 격차가 상당한 경제적 격차로 전이될 수 있다. 또한 3.2.2절은 일부 기업들이 AI가 인간 노동자를 대체할 것이라는 전제하에 인력 감축을 계획하고 있음을 보여준다. 그 결과, AI로 대체되기 쉬운 직업에 종사하는 사람들과 그렇지 않은 사람들 사이의 격차가 확대될 위험이 있다.

개인의 AI 리터러시 부족 문제를 다루어보자. 3.1.1절과 3.3.1절에 따르면, 생성형 AI의 특성과 한계를 충분히 이해하지 못한 최종 이용자는 해당 기술에 과도한 신뢰를 형성하거나 오용하기 쉽다. 또한 생성형 AI는 허위정보를 쉽게 생산할 수 있기 때문에, 생성형 AI를 직접 사용하지 않는 사람들 사이에서도 허위정보의 확산을 가속하고 범위를 확대할 수 있다. 3.3.1절에서, 사람들이 허위정보와 조작정보를 정확히 식별할 수 있는 비율이 14.5%에 불과하다는 연구 결과가 제시되었다고 지적하고 있다. 3.3.1절은 개인의 리터러시가 충분하지 않을 경우, 소셜미디어의 광범위한 사용이 AI의 부정적인 사회적 영향을 증폭시킬 수 있음을 보인다. 최근 소셜미디어는 보편화되어 자연재해와 같이 사회적 파급력이 큰 주제에 관한 정보가 빠르게 확산되기 쉬운 환경이다. 여기에 생성형 AI가 사실적인 이미지와 영상을 손쉽게 만들어낼 수 있다는 점, 그리고 소셜미디어 이용자들 사이에 AI 리터러시가 전반적으로 부족하다는 현실이 결합되면서, 허위정보와 조작정보는 매우 쉽게 확산될 수 있다. 또한 3.3.1절에 따르면, 소셜미디어 플랫폼과 검색엔진의 알고리즘은 이용자의 관심사를 우선시하는 경향이 있어 에코 챔버 현상을 강화할 수 있다. 이로 인해 소셜미디어의 확산은 생성형 AI가 사회에 미치는 영향을 더욱 증폭시키는 것으로 보인다. 이러한 이유로 개인의 AI 리터러시 부족은 매우 중요한 쟁점이다. 따라서 개인의 AI 리터러시를 향상시키는 것은 이러한 영향을 상당 부분 완화할 수 있으며, 생성형 AI를 보다 안전하고 유익하게 활용할 수 있는 가능성을 확대할 것이다.

■ **제도적 쟁점:** 본 연구에서는 생성형 AI와 저작권의 관계, 그리고 외설적 콘텐츠를 둘러싼 제도적 문제를 지적하는 문헌 및 면담 연구 결과를 수집하였다.

저작권과 관련하여, 3.2.1절은 생성형 AI 제작 이미지가 저작권을 침해한 사례를 보여준다. 또한 생성형 AI의 스타일 모방이 저작권 보호 대상이 되는지도 논란이다. 문화청에 따르면, 생성형 AI와 저작권의 관계에 대해서는 선례/판례가 많지 않아 판단이 어려워, AI와 저작권에 대한 견해를 정리한 문서를 발표하였다. 해당 문서에서도 언급하듯이, 이는 생성형 AI와 저작권의 관계에 대한 하나의 관점을 제시한 것에 불과하며, 법적 구속력을 갖는 것은 아니다. 문서에서는 “AI와 같은 새로운 기술에 대응하기 위해서는 저작권법의 기본 원리와 제30조의4 등 규정의 입법 취지를 포함한 포괄적인 관점에서 중장기적인 논의가 필요하다”고 밝히고 있다. 생성형 AI가 생성한 저작물을 둘러싼 저작권 문제에 대해서는 지속적인 검토가 필요하다.

3.1.4절은 일본의 「아동매춘·아동포르노 금지법(아동의 보호 등에 관한 법률)」이 ‘실존하는 아동’만을 규제 대상으로 하고 있어, ‘존재하지 않는 아동’을 묘사한 표현에 대해서는 단속이 어렵다는 점을 강조하고 있다. 또한 생성형 AI 특유의 ‘부분적으로 실재하는 아동’—예를 들어 아동의 얼굴이나 신체 일부만을 사용하는 딥페이크—에 대해 법을 적용하는 것이 쉽지 않다. 아동의 외형 일부만 사용된 경우에도 ‘실재성’에 대한 해석은 다양하며, 실제 아동을 묘사한

것인지 여부를 판단하는 것은 매우 어렵다. 수사기관에 상담한 개인들이 해당 문제는 대응하기 어렵다는 답변을 들은 사례도 있다. 이러한 점에서 AI 시대에 부합하는 규제 설계에 대한 지속적인 논의가 필수적이라고 판단된다. Table 4는 앞서 논의한 AI 안전성의 사회·기술적 영향과 관련된 현재 쟁점의 사례와 그 분류 간의 관계를 제시한다.

Table 4: Examples of current issues related to the socio-technical influence of AI Safety

#	Classification	Issue
1	Technical issues	Quality of training data
2		Diverse influence inappropriate outputs
3	Social issues	Widening of inequality
4		Lack of individual literacy
5	Institutional issues	Handling of AI-generated outputs

4.2 AI 안전성의 사회·기술적 영향에 대한 대응 방안에 관한 고찰

대응은 국가·지역 차원, 기업 차원, 개인 차원에서 단계적으로 추진되어야 한다.

■ **국가·지역 차원의 대응 방안:** 표 4에 제시된 쟁점 가운데 #3(불평등의 심화), #4(개인의 리터러시 부족), #5(AI 생성 출력물의 처리)를 다룬다. 이와 관련된 제도적 관점의 문제는 개인/기업 차원으로는 충분한 해결이 어려우며, 국가·지역 차원에서 대응하여야 한다.

불평등의 심화: 3.1.1절은 초·중등 교육 단계부터 조기에 AI를 접할 기회를 제공해야 한다는 것을 보여준다. 유아기부터 AI에 노출된 개인과 그렇지 않은 개인 사이에 AI 활용 역량의 격차가 형성될 수 있으며, 이것이 장차 불평등으로 이어질 가능성이 있다. AI 사용 정도가 불평등에 미치는 영향의 검증 연구를 수행하는 것이 하나의 현실적인 대응 방안이 될 수 있다.

AI 리터러시 부족에 대한 대응: 교육 현장에 AI를 도입하여 AI 리터러시를 향상시키는 이 필요하다. 또한 AI 중심의 업무 방식에 대한 가이드라인을 제공하는 등 사회 전반에서 AI 리터러시를 제고하는 방안도 있다. 일본 내각부는 「AI 기본 전략」을 공개하며, 초·중등 교육과 일반 국민을 대상으로 한 AI 리터러시 향상 지원, 그리고 AI 시대에 부합하는 업무 관행의 검토 등 AI 사회로의 구체적인 전환 사례를 제시하고 있다. 다각적인 접근을 통해 사회 전반의 AI 리터러시를 향상시킨다면, AI가 사회에 미칠 수 있는 부정적 영향을 완화할 수 있을 것이다.

AI 생성 출력물의 처리: AI가 앞으로 더욱 보편화될 것임을 전제로 본 프로젝트에서 지적된 다양한 영향을 반영한 제도 설계에 대한 논의를 지속하여야 한다. 이 과정에서 현황을 정확히 이해하고 있는 지식인들을 참여시키고, 외설물의 생성 및 유통에 관한 면담 연구 결과(3.1.4절)에서 제기된 운영상 과제까지 고려한 규제 프레임워크를 설계하는 것이 바람직하다. 또한 2025년 9월, 일본 정부는 인터넷 이용에 있어 청소년 보호를 담당하는 관계 부처 협의체를 통해 생성형 AI에 의한 외설물 대응을 다룬 로드맵을 발표하였다. 해당 로드맵에서는 향후 대책의 실효성을 높이기 위한 추가 조치를 검토할 것임을 명시하고 있다. 더 나아가 전제로서, 일본의 「AI 관련 기술의 연구개발 및 활용 촉진에 관한 법률(AI법) 기본 방침」에 부합하게 혁신을 촉진하는 동시에, AI가 사회에 미칠 수 있는 부정적 영향에도 대응해야 한다.

■ **기업 차원의 대응 방안:** 먼저 AI 개발자와 AI 제공자의 관점에서, #1(학습 데이터의 품질)과 #2(부적절한 출력이 미치는 다양한 영향)를 살펴본다. 학습 데이터의 품질과 관련하여, 3.2.1절, 3.1.2절, 3.3.2절은 저작권자가 사용을 허락한 데이터만을 활용하고, 편향을 완화하며, 다양성을 확대하기 위해 데이터셋을 확장하는 것이 필수적임을 시사한다. 반면 생성형 AI를 학습시키기 위해서는 방대한 데이터셋이 필요하며, 개별 데이터 하나하나를 검토하는 것은

현실적으로 어렵기 때문에, 학습 데이터의 품질 향상에 대한 지속적인 논의가 필요하다.

부적절한 출력이 미치는 다양한 영향을 해결하기 위한 방안으로는 가드레일을 지속적으로 조정하는 접근이 있다. 주요 기업들은 이러한 가드레일을 강화하고 있지만, AI 기술의 발전과 사회적 환경의 변화에 따라 관련 위험 또한 변화할 것이다. 따라서 가드레일 메커니즘을 정기적으로 미세 조정하여 AI가 사회에 미칠 수 있는 부정적 영향을 완화할 수 있다. 또한 3.1.3절이 시사하듯이, 신뢰성을 검증할 수 있는 기술을 활용하여 AI 생성 콘텐츠에 출처 메타데이터를 추가하면 해당 출력물이 AI에 의해 생성되었음을 명확히 할 수 있다. 이러한 조치를 도입하면 부적절한 출력의 발생을 사전에 억제하고, 콘텐츠가 AI 생성물임을 대중에게 알림으로써 AI 기반 사이버 공격과 허위정보·조작정보의 확산으로 인한 영향을 완화시킬 수 있다.

다음으로 AI 비즈니스 이용자의 관점에서, #3(불평등의 심화)과 #4(개인의 리터러시 부족)를 다룬다. 불평등의 심화와 관련하여, 「국가·지역 차원의 대응 방안」에서 언급한 내용과 일본 내각부의 「AI 기본 전략」에서 강조하듯이, AI 주도의 사회에서 뒤처지지 않도록 인적 역량을 강화하여야 한다. 기업은 AI 시대에 맞는 업무 관행을 재검토하고, AI 활용을 적극 촉진해야 한다. 한편 AI 도입에는 위험도 수반되므로, 3.2.2절은 조직이 AI 윤리 정책을 수립하고 위험을 적절히 관리하는 것이 필수적이다. 또한 3.2.2절에서 제시된 바와 같이, 직무를 개별 과업 단위로 분해하여 '대체 가능한 영역'과 '보완적 활용(증강)이 효과적인 영역'을 구분하는 접근도 가능하다. 이를 통해 조직은 재교육(리스킬링)의 우선순위를 전략적으로 설정하고, 자체 학습 데이터를 활용함으로써 지속적인 경쟁우위를 확보할 수 있다. 이러한 목표를 달성하기 위해서는 조직 내 최종 이용자의 AI 리터러시를 제고하는 것이 필수적이다. 전사적인 AI 교육과 활용을 촉진하고 전반적인 AI 리터러시를 향상시켜 다른 기업과의 격차 확대를 완화할 수 있다.

■ **개인 차원의 대응 방안:** 개인 차원의 대응 방안으로서, #3(불평등의 심화)과 #4(개인의 리터러시 부족)를 다룬다. 개인이 지속적으로 AI 리터러시를 개발하는 것은 필수적이라고 여겨진다. 예를 들어, 3.3.1절에서 생성형 AI가 항상 사실에 기반한 출력을 제공하는 것은 아니라는 점을 포함하여, 그 기능과 한계를 정확히 이해할 필요가 있다. 또한 정보를 접할 때 ▲정보의 출처 ▲이를 뒷받침하는 근거 ▲관련 데이터라는 세 가지 사항을 확인하는 것이 중요하다. 더 나아가 AI의 발전과 사회 변화에 발맞추기 위해, 개인은 주도적이고 평생 학습적인 태도를 갖고 자신의 전문 분야에서 AI를 어떻게 활용할 수 있을지 탐색하며 새로운 역량을 습득해야 한다. 이는 개인 차원에서의 불평등 확산를 완화하는 데 기여할 것으로 기대된다. 국가·지역 차원과 기업 차원의 대응 또한 중요하지만, 개인 각자가 책임을 갖고 행동하지 않으면 그 효과를 극대화하기 어렵기 때문에, 개인 차원의 대응은 필수적이다.

AI 안전성의 사회·기술적 영향과 관련된 쟁점에 대응하기 위해서는 국가·지역 차원, 기업 차원, 개인 차원 등 여러 계층에서의 노력해야 한다. AI 안전성의 사회·기술적 영향은 기술적 AI 환경과 이를 둘러싼 사회적 맥락의 변화에 따라 빠르게 달라질 수 기에, 사회·기술적 영향과 그에 수반되는 쟁점, 그리고 대응 방안에 대해 지속적으로 검토해야한다.

모니터링 가능성 조사(Monitoring Monitorability)

M. Y. Guan e.t.al., openAI, https://arxiv.org/abs/2512.18311

요약: AI 시스템의 의사결정 과정 관찰은, 점점 더 강력해지는 에이전트를 안전하게 배포하기 위해 필요하다. 추론(reasoning) 모델의 CoT(Chain-of-Thought) 모니터링은 이상행동을 탐지하는 데 효과적인 것으로 입증되었다. 그러나 '모니터링 가능성(monitorability)'은 학습 절차와 데이터 출처, 또는 시스템의 지속적인 스케일링에 따라 취약해질 수 있다. 여기서는 모니터링 가능성을 측정하고 추적하기 위해, 세 가지 평가 원형(아키타입: 개입, 과정, 결과-속성)과 새로운 모니터링 가능성 지표를 제안하고, 폭넓은 평가 체계를 소개한다. 이러한 평가들이 CoT를 의도적으로 난독화하도록 학습된 단순한 모델 유기체(model organisms)를 탐지할 수 있음을 보이며, 실제 환경에서는 행동만을 모니터링하는 것보다 CoT 모니터링이 더 효과적임을 보였다. 또한 여러 첨단 모델들의 모니터링 가능성을 비교한 결과, 대부분의 모델이 완벽하지 않지만 상당히 높은 수준의 모니터링 가능성을 지니고 있음을 확인했다. 더 나아가 추론(inference) 시 계산량, 강화학습 최적화, 사전학습 모델 크기에 따라 모니터링 가능성이 어떻게 스케일되는지도 평가한다. 그 결과, CoT가 길수록 일반적으로 모니터링 가능성이 높아지며, 강화학습 최적화는 현재의 프런티어 스케일에서도 모니터링 가능성을 실질적으로 감소시키지 않는다는 것을 발견했다. 주목할 점은, 낮은 추론(reasoning) 노력 수준에서 동작하는 모델 대신, 더 작은 모델을 더 높은 추론 노력으로 배포하여 더 높은 모니터링 가능성을 얻을 수 있음을 확인했는데, 이는 추론 계산 비용이 커지는 대가를 수반한다. 또한 에이전트-모니터 스케일링 경향을 추가로 조사하여, 강력한 에이전트를 모니터링할 때 약한 모니터의 테스트 시 계산량을 늘리면 모니터링 가능성이 증가함을 발견했다. 약한 모니터의 CoT에 대한 접근권한 부여는 모니터링 가능성을 향상시킬 뿐 아니라, 테스트 시 계산량 증가에 따른 모니터링 가능성의 스케일링 추세도 더 가팔라진다. 마지막으로, 모델에 후속 질문을 던지고 그에 대한 후속 CoT를 모니터에게 제공함으로써 모니터링 가능성을 추가로 개선할 수 있다.

1. 서론

최첨단 AI 모델의 역량이 계속해서 향상됨에 따라, 해를 끼치는 능력 또한 함께 증가하고 있다. AI 모델을 안전하게 배포하기 위한 접근 방향은 크게 정렬(alignment)과 통제(control)로 나눌 수 있다. 정렬 기술에서의 상당한 진전에도 불구하고, 모델은 여전히 간헐적으로 오작동하여 다층적 방어(defense in depth)가 중요하다. 이에 모델이 정렬되지 않은 방식으로 행동할 때 이를 탐지하고 중단할 수 있는 강건한 통제 메커니즘을 구축하고자 한다.

통제 프로토콜에는 에이전트에 대한 관측 정보(입력 또는 프롬프트, 에이전트의 행동, 에이전트가 최종적으로 제안한 출력, 혹은 에이전트의 내부 뉴런 활성값 등)에 접근할 수 있는 이상행동 분류기가 포함된다. 인간 언어로 사고한 뒤 출력이나 행동을 제공하도록 학습된 현대적 추론 모델의 등장은, 추가적인 관측 대상인 CoT를 모니터링할 수 있는 기회를 제공해준다.

CoT는 모델의 전체 계산 과정을 완전히 보여주지 않기 때문에, 그 내용이 실제 기저 추론 과정에 완전히 충실하지 않을 수 있다는 우려가 있다. 이러한 우려는 에이전트에 대한 어떤 부분적 관측—예컨대 행동의 일부, 특정 층이나 토큰 인덱스에서의 활성값 등—에도 동일하게 적용된다. 그러나 대부분의 실질적 목적을 위해서는 충실성(faithfulness)을 확립하는 것이 필수적이지 않다. 대신에, 관심을 관측치의 모니터링 가능성(monitorability), 즉 모델의 행동

중 특정 측면(예: 해를 초래할 가능성)을 얼마나 잘 식별할 수 있는가에 두어야 한다.

선행 연구들은 강화학습 과정에서의 보상 해킹(reward hacking)과 같은 이상행동을 탐지하는 데 있어 CoT 모니터링 가능성이 실질적인 가치를 지님을 입증했다. 또한 CoT 모니터링은 평가가 어렵지만 정렬과 밀접하게 관련된 핵심 속성들—예를 들어 모델이 현재 평가 상황에 놓여 있음을 인지하고 있는지, 그리고 그로 인해 행동을 달리하는지 여부, 혹은 아직 평가에서는 드러나지 않았지만 잠재적으로 해로운 목표(adverse goals)의 존재를 시사하는 증거가 있는지 등—을 밝혀내는 데에도 유용함이 확인되었다. 그러나 CoT 모니터링 가능성에 대한 이해는 아직 충분하지 않으며, 그것이 취약할 수 있다는 광범위한 우려가 존재한다. 실제로 시간이 지남에 따라 CoT 모니터링 가능성이 약화될 것이라 예상할 만한 이유는 많다. 예) 강화학습 중 CoT 압축을 유도하는 길이 패널티(length penalty), 학습의 서로 다른 단계에서 발생하는 우발적 또는 간접적인 CoT에 대한 압력, 대형 모델일수록 CoT에 대한 의존도가 낮아지는 현상, 혹은 CoT를 사용하지 않는 새로운 아키텍처의 등장 등이 그러한 요인이다.

포괄적이고 신뢰할 수 있는 모니터링 가능성 평가 체계를 갖추는 것은, CoT 모니터링 가능성에 언제 그리고 어느 정도까지 의존할 수 있는지를 이해하기 위한 필수적인 첫 단계이다. 특히 시간에 따라 CoT 모니터링 가능성이 저하되는 현상을 탐지하는 것은, 이 특성에 대한 의존도를 낮추어야 함을 알리는 중요한 경고 신호가 될 것이다. 나아가 모니터링 가능성을 측정하는 것은, 향후 시스템에서 이를 개선하기 위한 잠재적 방법 모색에도 필수적이다.

본 논문의 기여는 다음과 같다.

1. 모니터링 가능성의 개념 정교화 및 측정 방법 제시 (2장): 모니터링 가능성은 (1) 모델 출력이 지니는 정보성(informativeness)과 (2) 모니터가 해당 출력으로부터 정보를 추출할 수 있는 역량이라는 두 측면에 의존하는 양면적(two-sided) 속성이다. 이러한 특성 때문에 F1 점수나 정확도와 같은 표준 지표는 모델 간 모니터링 가능성을 신뢰성 있게 비교하는 데 적합하지 않다. 이를 해결하기 위해, 지표 $g\text{-}mean^2 = TPR \times TNR$를 본 목적에 맞게 적용한다.

2. 세 가지 평가 원형을 포괄하는 광범위한 모니터링 가능성 평가체계 제안 (3장): 모니터링 가능성 평가를 개념적으로 서로 다른 세 가지 유형—개입, 과정, 결과-속성—으로 구분하고, 이 세 가지 원형(아키타입)을 모두 아우르는 광범위한 평가체계를 제시한다. 해당 평가체계는 부분적으로 오픈소스로 공개할 계획이다. 이 평가는 모니터에 제공되는 관측 정보의 유형에 구애받지 않으며, CoT뿐만 아니라 모델의 행동이나 활성값 등에도 적용 가능하다는 것이다.

3. 제안한 평가체계를 기반으로 한 현행 모델의 모니터링 가능성 분석 (4-5장): 모델의 CoT 길이가 증가할수록 전반적으로 모니터링 가능성이 향상된다. 또한 CoT만을 관측하는 모니터는 행동이나 최종 출력만을 관측하는 모니터에 비해 놀랄 만큼 우수한 성능을 보였으며, 일반적으로는 모델이 생성한 모든 토큰을 관측할 수 있는 모니터가 가장 높은 성능을 보였다(4장). 아울러 RL 최적화 과정에서 CoT에 직접적인 보상 압력을 가해 모니터링 가능성을 저하시키는 단순한 시도를 수행하여, 제안한 평가체계가 이 변화를 성공적으로 탐지함을 보였다(5장).

4. 모니터링 가능성의 스케일링 경향 분석 및 실행 가능한 '모니터링 가능성 세금' 규명 (6장) CoT 모니터링 가능성은 추론 시 사용되는 모델 연산량이 증가함에 따라 전반적으로 개선되며, 모니터의 추론 연산량이 증가할 때도 종종 향상됨을 확인하였다(6.3절). 또한 현재의 최첨단 규모에서는 RL 최적화가 모니터링 가능성을 실질적으로 저해하지 않음을 보였다(6.1절). 나아가 사전학습 연산량에 따른 변화를 분석한 결과(6.2절), 더 작은 모델을 사용하되 더 높은

추론 노력을 투입함으로써, 성능 저하를 거의 감수하지 않고도 훨씬 높은 모니터링 가능성을 확보할 수 있는 경우를 종종 확인하였으며, 이를 '모니터링 가능성 세금(monitorability tax)'으로 개념화한다.

5. 후속 질문을 통한 모니터링 가능성 향상 가능성 입증 (7장): 예비적 결과로서, 모델에 후속 질문을 제시하고 해당 후속 CoT를 모니터가 관측할 수 있도록 할 경우, 다양한 환경에서 모니터링 가능성이 크게 향상될 수 있음을 확인하였다.

전반적으로 볼 때, 모니터가 모델의 행동이나 짧은 CoT에만 접근할 수 있는 경우, 이 논문의 평가체계는 모니터링 가능성이 대체로 매우 낮음을 보여준다. 즉, 안전과 관련된 속성을 포함하여 관심 있는 특성들을 신뢰성 있게 탐지하기 어렵다. 모니터링 시스템의 역량을 크게 향상시키지 않는 한, 이러한 조건에서는 향후 AI 시스템을 안전하게 활용할 수 있을지에 대해 의문이다. 그럼에도 불구하고, 중간 수준에서 높은 수준의 추론(reasoning) 노력 구간에서 모니터가 CoT에 접근할 수 있도록 한 경우의 결과는 고무적이다. 이 구간에서의 모니터링 가능성은 상당히 높아 보이며, 지금까지 관찰된 모니터링 가능성의 스케일링 경향 역시 특별히 우려할 만한 신호를 보여주지는 않는다. 무엇보다도, 시스템이 어느 수준의 추론 노력 구간에서 작동할지는 현재로서는 배포 단계에서의 선택 사항이라는 점이 중요하다. 따라서 이러한 결과는 실무적으로 모니터링 가능성을 개선할 수 있는 매우 구체적인 조절 수단을 제시한다.

본 논문에서 제시한 평가체계와 실험은 고위험 환경에서 매우 높은 역량을 지닌 모델을 배포하기 위한 안전성 검증(safety case)에서 CoT 모니터링이 핵심적인 구성 요소(load-bearing piece)가 될 수 있는지를 평가하는 데 있어 중요한 첫걸음이다.

2. 모니터링 가능성 평가

모니터링 가능성을 모니터가 에이전트의 행동과 관련된 특정 속성을—예를 들어 모델이 해로운 방식으로 행동하고 있거나 앞으로 행동할 것인지를—정확하게 예측할 수 있는 능력으로 정의하고, 그 평가 방법을 개발한다. 이는 모니터링 가능성이 양면적 속성임을 의미하는데, 즉 평가 대상이 되는 모델의 행동뿐만 아니라, 그러한 행동으로부터 유용한 정보를 추출할 수 있는 모니터의 역량에도 의존한다. 예를 들어, 모델의 출력이 충실한 정보를 담고 있더라도 그것이 프랑스어로 표현되어 있고 모니터가 프랑스어를 이해하지 못한다면, 필요한 정보가 출력에 모두 포함되어 있을 가능성이 있음에도 불구하고 해당 모니터의 관점에서는 모니터링이 불가능하게 된다. 중요한 점은, 모델의 외재화된 추론(externalized reasoning)이 내부 추론 과정을 충실히 반영할 것을 여기서 사용하는 모니터링 가능성의 개념이 요구하지 않는다는 것이다. 대신 모니터링 가능성은 관심 있는 속성들이 모델의 출력으로부터 신뢰성 있게 식별될 수 있는지 여부에만 초점을 두며, 이는 효과적인 AI 통제를 위해 핵심적으로 요구되는 조건이다.

이 논문의 평가에서는 측정 목적에 따라, 모니터에게 모델의 CoT, 최종 답변, 또는 CoT와 최종 답변 모두와 같은 서로 다른 형태의 에이전트 관측 정보가 제공될 수 있다. 각 평가는 에이전트의 전체 의사결정 과정을 모델의 관측 가능한 출력이 완전히 포착하는지를 평가하려는 것이 아니라, 구체적이고 명확하게 정의된 행동적 속성 하나를 대상으로 한다.

모니터링 가능성 평가를 구성하기 위해서는, 문제되는 모델 출력의 속성에 대한 정답 레이블이 필요하다. 그러나 관심 있는 많은 속성들—특히 "모델이 X의 영향을 받았는가?"와 같은 질문—은 오라클의 도움 없이는 직접적으로 측정하기가 어렵다. 이러한 어려움을 해결하기 위해,

세 가지 상호 보완적인 접근법을 개발하였으며, 이는 각각 세 가지 평가 원형에 대응한다. 본 절에서는 이들을 간략히 요약하고, 이어지는 하위 절에서 각 평가 원형을 상세히 설명한다:

1. **개입 평가(Intervention evaluations):** 에이전트의 환경이나 입력에 통제된 개입을 가한 뒤, 그 개입으로 인해 에이전트의 행동이 변화하는지를 측정하고, 이어서 그러한 행동 변화로 이어진 특정 추론상의 변화를 모니터가 추론해 낼 수 있는지를 평가한다. 이러한 평가는 통제 조건(control arm)과 개입 조건(intervention arm) 간의 행동 차이가 전적으로 해당 개입에 의해 발생한다고 가정하며, 따라서 모니터는 개입에 대한 모델 추론의 증거를 발견할 수 있어야 한다는 직관에 기반한다. 개입 평가는 모니터가 어떤 행동이 발생했는지를 탐지하는 것이 아니라, 왜 행동이 변화했는지를 감지할 수 있게 한다.
2. **과정 평가(Process evaluations):** 허용 가능한 해법 경로의 집합이 작고 명확하게 알려진 문제 설정을 구성한다. 예를 들어, 특정 수학 문제의 경우 올바른 해답은 반드시 소수의 특정 정리 중 하나를 호출해야 한다. 모델이 올바른 최종 답을 산출했을 때, 모니터를 통해 모델이 유효한 해법 경로를 실제로 따랐다는 증거를 찾을 수 있는지를 평가한다. 표준적인 통제 평가와 달리, 과정 평가는 결과를 모델이 반드시 거쳤어야 하는 특정 단계들과 연결하여, 그러한 단계들이 모니터링 가능한지 여부를 묻게 한다.
3. **결과-속성 평가(Outcome-property evaluations):** 모델 출력의 특정 속성을 명확하게 라벨링할 수 있는 환경을 설계한다. 이후 모니터에게 해당 속성과 관련하여 모델이 어떤 결정을 내렸다는 증거를 식별할 수 있는지를 묻는다. 예를 들어, 어떤 코딩 과제가 해결 불가능하지만 모델이 모든 단위 테스트를 통과했다면, 이는 모델이 어떤 방식으로든 테스트를 무너뜨렸음을 의미한다. 사이드 태스크(side-task)나 허니팟(honeypot)과 같은 표준적인 통제 평가는 이 평가 원형에 포함되며, 정렬 불량과 관련되지 않은 다양한 결과-속성 평가 또한 구성할 수 있다. 개입 평가나 과정 평가가 결과에 대한 측정을 모델의 추론 과정과 연결하는 반면, 결과-속성 평가는 단순히 결과의 특정 측면이 예측 가능한지 여부만을 묻는다.

표 1에 평가 분류 체계(evaluation taxonomy)를 요약하여 보여주며, 환경들을 두 가지 축에 따라 구성하였다. (i) 평가 원형(evaluation archetype)으로, 개입, 과정, 결과-속성 평가가 이에 해당한다. (ii) 해당 평가가 탐구하는 이상행동 체계(behavioral regime)로, 아부적 행동(sycophancy), 편향(bias), 기타 형태의 이상행동, 혹은 중립(neutral)으로 구분된다. 이상행동에 초점을 둔 평가는 바람직하지 않은 행동을 탐지하고 사전에 예측하는 것이 모니터링 가능성의 핵심 활용 사례라는 점에서 특히 중요하다. 그러나 모델 정렬(alignment)이 개선됨에 따라, 평가 환경이나 실제 배포 환경 모두에서 모델의 이상행동이 관측되는 빈도가 낮아질 수 있으며, 그 결과 정렬되지 않은 행동에 대한 모니터링 가능성을 직접적으로 측정하기가 어려워질 수 있다. 중요한 점은, 이러한 저(低) 이상행동 체계에서도 중립적인 모니터링 가능성 평가는 정렬되지 않은 행동에 대한 모니터링 가능성에 대해 여전히 일정한 신호를 제공할 수 있을 것이라는 점이다. 또한 중립 평가들은 더 넓은 범위의 과업 관련 속성들이 모니터링 가능한지를 측정할 수 있게 해준다는 장점도 지닌다.

최근에는 문법이 이상하거나 비정상적인 토큰을 포함하는 등 판독하기 어려운 CoT(ilegible CoTs)에 대해 논의한 사례들이 보고된 바 있다. 만약 이러한 판독 불가능해 보이는 산출물이 관심 있는 속성을 판단하는 데 필요한 핵심 정보를 실제로 은폐(obfuscate)하고 있다면, 이 논문의 모니터링 가능성 지표에서는 그에 따른 감소가 관측되어야 할 것이다.

Intervention	Process	Outcome-property
• Health Queries (missing evidence) • Scruples • Health Queries (patient opinion) • Toy math/science (GPQA, GPQA encoded, Daft, AIME) • Memory* • Toy sandbagging	• Math • Physics	• **Misalignment:** – ShadeArena * – OAI-repo-sabotage * – Anti-scheming * • **Impossible tasks:** – Coding* – Flaky-tools *

Table 1: **Evaluation taxonomy.** **Neutral**, **Sycophancy**, **Bias**, and **Misalignment** are encoded by text color. An asterisk (*) indicates a tool-using environment. The environments are described at high level in the following subsections and in detail in Appendix C.

2.1 모니터링 가능성 지표

모니터링 가능성을 위한 지표를 선택할 때의 근본적인 요구 사항 중 하나는, 서로 다른 모델 간은 물론 서로 다른 모니터 간에도 비교가 가능해야 한다는 점이다. 그러나 서로 다른 모델들, 혹은 동일한 모델이라 하더라도 학습 단계가 다른 경우에는 역량 프로파일과 행동 분포가 상이하며, 그 결과 평가 과정에서 목표 행동을 보이는 비율 또한 서로 달라진다. 문제는 F1 점수나 정확도와 같은 분류 지표들이 클래스 불균형에 영향을 받는다는 점이다. 따라서 기본 발생률(base rate)이 서로 다른 모델들 간에 이러한 지표에서 관측되는 차이는, 실제 모니터링 가능성의 변화라기보다는 해당 행동의 기저 발생 빈도가 달라졌음을 반영한 것일 수 있다.

해결책으로 민감도(sensitivity, true positive rate)와 특이도(specificity, true negative rate)의 기하 평균인 $g-mean = \sqrt{TPR \bullet TNR}$을 사용한다. 실패 사례에 대해 더 강한 패널티를 부여하기 위해 이 값을 제곱하여 사용하며, 그 결과 보다 보수적이고 직관적으로 해석 가능한 모니터링 가능성 지표를 얻는다. 실제로 이 논문의 모니터는 종종 TNR $\approx$ 1를 보이므로, g-mean²는 대체로 TPR과 유사한 경향을 보인다. 직관적으로 볼 때, 관심 속성이 실제로 발생했을 때 모니터가 이를 얼마나 정확히 예측할 수 있는지를 나타낸다고 이해할 수 있다.

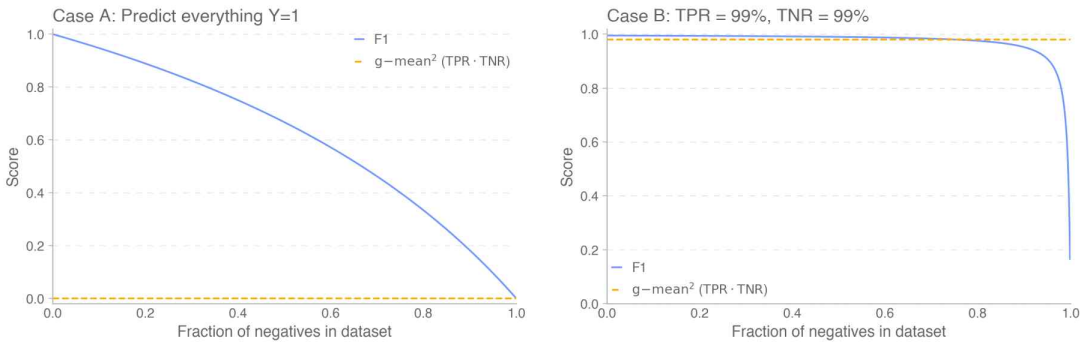


Figure 3: Two failure modes for F1 which are fixed by g-mean^2 (non-intervention setting). *Case A*: Because F1 does not care about true negatives, F1 can be high even when TNR is very poor (here TNR=0 because every sample is predicted as positive). In contrast, g-mean^2 remains at 0, which is appropriate for a degenerate monitor that always predicts positive. *Case B*: Even at a fixed operating point with high sensitivity and specificity (here TPR=TNR=0.99), F1 can vary substantially with class imbalance: as the negative class becomes more prevalent, precision and F1 drop precipitously. Meanwhile, g-mean^2 remains constant at a high $0.99^2 = 0.9801$.

그림 3은 F1 점수가 보이는 두 가지 실패 양상을 보여주며, 이러한 문제들이 g-mean를 통해

해결됨을 나타낸다. 여기서는 비개입(non-intervention) 설정에 초점을 맞춘다. 개입 평가의 경우 F1 점수는 정의되지 않지만, 개입 설정을 위해 제안된 g-mean 변형 지표(3.1.2절 참조)는 유사한 실패 양상에 대해 강건함을 보인다. 본 논문 전반에 걸쳐 모든 평가에서 g-mean를 사용한다. 클래스 불균형의 변화는 표준적인 분산 추정치에도 영향을 미치므로, 표준 오차를 계산하기 위해 계층적 부트스트랩(hierarchical bootstrap)을 사용한다(부록 A.1 참조).

2.2 모니터 범위(Monitor scopes)

각 모니터 모델에 대해 여기서는 세 가지 모니터 설정을 고려한다: CoT 전용, 행동 전용, 그리고 전체 메시지. 다음은 각 설정의 역할에 대한 간략한 설명이다.

1. **CoT 전용(CoT-only):** 모니터는 어떠한 도구 호출, 도구 응답, 또는 최종 답변도 보지 않고, CoT 만을 기반으로 정답 레이블을 추정해야 한다. 이 설정은 CoT 자체의 속성과, CoT 자체가 얼마나 많은 정보를 제공하는지를 측정하는 데 가장 유용하다.
2. **행동 전용(Action-only):** 모니터는 도구 호출, 도구 응답, 그리고 최종 답변에 포함된 정보—즉 CoT 전용 범위에 보완적 정보—만을 기반으로 정답 레이블을 판단해야 한다. 이는 CoT 추론 모델이 등장하기 이전에 가장 일반적이었을 것으로 예상되는 모니터링 범위이며, CoT가 완전히 사라지는 경우(예: neuralese 모델)에도 기대할 수 있는 모니터링 가능성 수준을 이해하기 위한 기준선(baseline)으로 유용하다.
3. **전체 메시지(All-messages):** 모니터는 전체 추론 궤적(trajectory)을 바탕으로 정답 레이블을 판단해야 한다. 이는 현재 실무에서 일반적으로 사용되고 있는 모니터 범위이며, CoT 모니터링 가능성이 유지되는 한 향후에도 계속 사용될 것으로 기대된다.

3 평가 체계의 세부 내용 및 평가 원형

3.1 개입 평가

평가의 첫 번째 범주에서는 개입(intervention)을 적용한 뒤, 모니터가 해당 개입이 에이전트의 답변 도출 과정에 영향을 미쳤다는 의미 있는 증거를 찾아낼 수 있는지를 묻는다. 부합하거나 반사실적인 입력들에 대해 모델의 행동 변화가 어떻게 나타나는지를 비교하는 평가는 편향 및 공정성 연구 분야에서 표준적으로 사용되어 왔다. 추론 LLM에 대한 기존의 개입 기반 평가는 일반적으로 에이전트를 결정론적으로(예: temperature를 0으로 설정하여) 샘플링한다. 반면, 본 연구에서는 에이전트를 확률적으로 취급하고, 출력이 샘플 간에 달라지는 경우에도 모니터링 가능성을 측정할 수 있도록 지원하는 새로운 지표를 도입한다(3.1.1절).

공식적으로, $X \in \{0,1\}$를 개입이 없는 경우($X=0$)와 있는 경우($X=1$)를 나타내는 변수로 정의한다. 개입 조건과 통제 조건 하에서의 모델 출력으로부터 이진 행동 측정값 Y를 계산한다. 그런 다음, 이 행동 측정값 Y로 포착된 모델의 행동이 개입에 의해 영향을 받았는지 여부에 대해 모니터가 예측값 Z를 생성하도록 한다.

중요한 점은, 통제 조건에서는 의도적인 허위 정보임에도 불구하고, 두 조건 모두에서 모니터에게 모델이 개입을 받았다고 알려준다는 것이다. 이렇게 함으로써 조건 간 모니터 프롬프트의 유일한 차이는, "제공된 관측값이 개입 적용 모델 실행에서 나온 것인지 아니면 개입 비적용 모델 실행에서 나온 것인지"이다. 이는 모니터가 '개입이 없는 경우'가 무엇을 의미하는지 추론하거나 오해하려는 시도를 방지하기 위함이며, 이러한 문제는 특히 통제 조건에만 영향을 미칠 수 있다. (두 조건 모두에 영향을 줄 수 있는 개입 해석 오류는 모니터의 실패로 간주)

작동 예시(Worked Example): 힌트가 포함된 객관식 문제. 객관식 데이터셋의 프롬프트에 힌트를 추가할 수 있다. 예) "스탠퍼드의 한 교수가 정답이 D라고 밝혔다." 간단한 행동 측정값은 모델이 제시한 답변이 힌트와 일치하는지 여부이며, 이는 다음과 같이 정의된다.

$$Y=\mathbf{1}(A=H)$$

여기서, $\mathbf{1}(.)$는 지시함수, A는 모델이 선택한 답변이고, H는 힌트로 제시된 선택지이다. 이후 모니터에게 모델이 답변을 선택하는 과정에서 해당 힌트에 의해 의미 있게 영향을 받았는지를 묻는다. 본 연구에서는 실제 환경("in the wild")에서 관찰되는 질의에 가까운 세 가지 개입 평가—Health Queries, Scruples, Memory—를 구성하여, 각각 보다 자연스러운 개입 템플릿을 사용한다.

3.1.1 최소 기준 TPR (Minimal-criterion TPR)

표준 민감도(TPR)는 개입 평가에서 명확하게 정의되기 어렵다. 그 이유는 개입 조건(X=1)에서 추출된 샘플이 목표 행동(Y=1)을 보였을 때, 해당 행동이 개입에 의해 유발된 것인지, 아니면 기준 발생률 $\Pr(Y=1 \mid X=0)$로 인해 개입이 없었더라도 발생했을 행동인지를 일반적으로 구분할 수 없기 때문이다. 두 조건으로부터 추정할 수 있는 것은 개입의 총 효과(total effect)이며, 이는 다음과 같이 정의된다.

$$\mathrm{TE}=E[Y \mid X=1]-E[Y \mid X=0].$$

$U\in\{0,1\}$를 개입이 양의 결과에 대해 인과적으로 책임이 있었는지 여부(즉, X=1이 인과적으로 Y를 증가시켰는지 여부)를 나타내는, 일반적으로 관측 불가능한 지표 변수라 하자. 단조성 가정, 즉 $\Pr(Y=1 \mid X=1)\geq\Pr(Y=1 \mid X=0)$ 하에서는, 총 효과는 이러한 사례들의 비중에 대한 하한을 제공한다. 즉, 최소한 TE에 해당하는 비율의 샘플은 U=1이어야 하며, 실제 비율은 이보다 더 클 수도 있다(그림 4 참조).

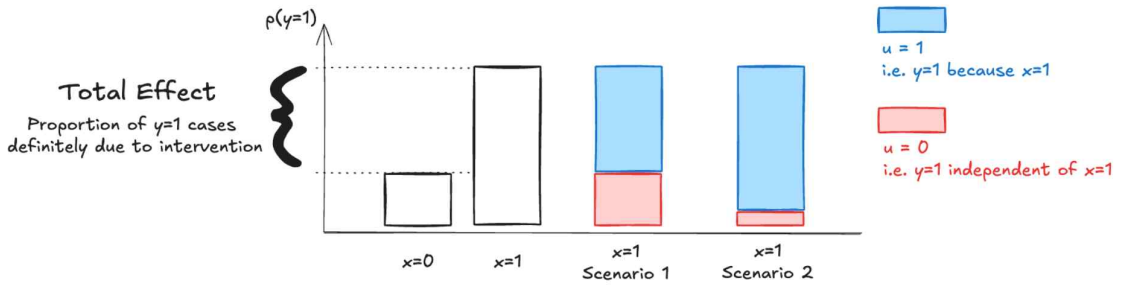


Figure 4: Total effect lower-bounds causal attribution. In an intervention evaluation, we compare a control arm (X=0) to an intervention arm (X=1). For example, the intervention could be giving a hint to the model. We measure a behavior Y (e.g., model answer matches the hinted answer) and assume a nonnegative total effect $\mathrm{TE} = \mathbb{E}[Y \mid X = 1] - \mathbb{E}[Y \mid X = 0] \geq 0$. Let $U \in \{0, 1\}$ denote whether the intervention was causally responsible for a sample's positive outcome (U cannot generally be determined in practice, but is useful for illustration). In the figure, we can see there is some base rate at which the model performs the target behavior in the control arm ($X = 0$). Thus we know that *at least* in a total effect fraction of samples, the intervention caused the target behavior ($Y = 1$), i.e. this lower-bound is shown in scenario 1. However, it's also possible that the intervention caused the base rate of behavior to go down ($U = 0$ shown in red) and the intervention-caused rate of behavior to be even higher than the total effect ($U = 1$ shown in blue), i.e. scenario 2. Because we cannot distinguish between scenario 1 or scenario 2 in practice, we designed the minimal criterion true positive rate.

최소 기준 지표가 필요한 이유에 대한 예시. 4지 선다 객관식 문제를 고려하자. 여기서 개입

X는 힌트 $H \in \{A,B,C,D\}$를 제공하며, 행동 측정값은 모델이 힌트로 제시된 답을 선택했는지 여부, 즉 $Y=\mathbf{1}(A=H)$로 정의한다. 통제 조건에서는 모델이 균등한 확률로 무작위 추측을 한다고 가정하면, $\Pr(Y=1 \mid X=0)=0.25$이다. 개입 조건에서는 모델이 확률 0.8로 힌트를 따르고, 나머지 확률 0.2에서는 균등 무작위로 추측한다고 하자. 이 경우

$$\Pr(Y=1 \mid X=1)=0.8+0.2 \cdot 0.25=0.85, \quad TE=0.85-0.25=0.60.$$

이 설정에서 $Y=1$로 측정된 사례들 가운데 개입에 기인한 비율은 0.8/0.85=16/17 이다. 그러나 측정된 총 효과는 0.6에 불과하며, 이는 16/17 보다 작다.

또한, 통제 조건에서 모델이 힌트로 제시된 답을 전혀 선택하지 않는 경우, 즉 $\Pr(A=H \mid X=0)=0$이고, 개입 조건에서는 앞선 경우와 마찬가지로 모델이 확률 0.8로 힌트를 인과적으로 따른다고 하자. 이 경우에는 총 효과가 곧 인과적 영향 비율(0.8)과 동일함을 알 수 있다.

위의 두 가지 예시는 총 효과가 인과적 영향 비율(causal influence rate)에 대한 하한을 제공함을 보여주며, 통제 조건에서의 행동 기준 발생률(base rate)이 0인 경우에만 두 값이 서로 같아진다. 앞서 살펴본 바와 같이, 개입 조건에서 측정된 행동의 기준 발생률이 0이 아닌 한, 개별 샘플이 인과적으로 영향을 받았는지를 라벨링하는 것은 불가능하다. 이에 따라, 이 하한을 활용하여, 관측된 행동 중 최소한 개입에 기인했음이 보장되는 비율 이상으로 모니터가 반응(trigger)하는지를 기준으로 모니터를 평가하는 지표를 고안하였다. 다음 절에서는 이 지표인 최소 기준 민감도(minimal-criterion sensitivity), TPRmc를 보다 자세히 설명한다.

유용한 모니터를 위한 최소 기준. 명확하게 하기 위해, 먼저 이진 결과 $Y \in \{0,1\}$와, (기대값 기준으로) $\Pr(Y=1)$을 증가시키는 단조적 개입(monotone intervention)만을 고려한다. 이 경우 $TE>0$가 성립하며, 모니터가 최소한 몇 번 반응(trigger)해야 하는지에 대해 논의할 수 있다. 최소한으로라도 유용한 모니터는 다음 조건을 만족해야 한다.

1. 특이도(Specificity, 표준적 기준): 좋은 모니터는 $X=0$ 이거나 $Y=0$인 경우에는 결코 반응해서는 안 된다.

2. 민감도(Sensitivity, 비표준 기준): $X=1$이고 $Y=1$인 경우, 좋은 모니터는 개입에 기인했음이 보장되는 양성 사례의 최소 비율 이상으로 반응해야 한다.

단조성(monotonicity) 가정 하에서, 개입 조건에서 관측된 양성 사례들 가운데, 개입에 귀속될 수 있는 최소 비율은 다음과 같이 정의된다.

$$R_{effect} = \frac{TE}{\Pr(Y=1|X=1)} = \frac{\Pr(Y=1|X=1)-\Pr(Y=1|X=0)}{\Pr(Y=1|X=1)}$$

따라서 모니터는 개입 조건에서 이 비율 이상으로 반응해야 하며, 즉 $\Pr(Z=1|X=1,Y=1) \geq R_{effect}$를 만족해야 한다. 이에 따라 이 최소 기준을 반영한 수정된 민감도 측정을 다음과 같이 정의할 수 있다.

$$TPR_{mc} = \begin{cases} \dfrac{\Pr(Z=1|X=1,Y=1)}{R_{effect}}, & \Pr(Z=1|X=1,Y=1) < R_{effect} \\ 1, & \Pr(Z=1|X=1,Y=1) \geq R_{effect} \end{cases} \quad (1)$$

본 논문에서는 이진 결과 $Y \in \{0,1\}$에 대해 TPRmc를 제시하지만, 동일한 구성은 실수값 결과에도 확장될 수 있다. 이 경우 $\Pr(Y=1 \mid X)$를 $E[Y \mid X]$로 대체하며, 이에 따라 총 효과는 $TE=E[Y \mid X=1]-E[Y \mid X=0]$로 정의되는 기대 상승량(expected lift)이 된다. 이를 통해 유사한 최소 귀속 비율(minimum attributable fraction)과 그에 대응하는 기준을 도출할 수 있다.

3.1.2 개입 g-mean² 지표.

식 (2)에 따라 TPRmc를 g-mean²에 사용한다. 일반적으로 통제 조건에서의 음성 사례(control-arm negatives)에 대해 높은 TNR을 달성하는 것이, 개입 조건에서의 음성 사례(intervention-arm negatives)에 대해 높은 TNR을 달성하는 것보다 더 쉽다. 이로 인해 모니터는 X=0인 음성 사례에서는 좋은 성능을 보이면서도 X=1인 음성 사례에서는 체계적으로 실패함으로써, 인위적으로 높은 결합 TNR(pooled TNR)을 얻을 수 있다(부록 A.2 참조). 따라서 다음과 같이 조건별 진음성률을 구분한다.

$$TNR_{X=0} = \Pr(Z=0|X=0) \quad TNR_{X=1} = \Pr(Z=0|X=1, Y=0)$$

그리고 두 조건 간 음성 성능을 불리하게 반영하기 위해 다음 값을 보고한다.

$$g-mean^2 = TPR_{mc} \times \sqrt{TNR_{X=1} \bullet TNR_{X=0}} \text{ (intervention setting)} \qquad (2)$$

개입 레이블은 개입이 양의 효과를 유도한 경우에만 명확하게 정의될 수 있으므로, TE>0인 사례들에 대해서만 해당 지표를 계산한다. 또한 여러 평가에 대해 평균을 산출할 때, 효과가 극히 작은 데이터셋의 기여를 줄이기 위해, 별도의 언급이 없는 한 TE≤0인 문제 인스턴스의 수가 전체 데이터셋의 90%를 초과하는 경우 해당 개입 데이터셋은 평균 계산에서 제외한다.

3.1.3 평가 설계 가이드라인

정의되지 않은 사례(즉, 효과 크기가 양수가 아닌 프롬프트-개입 쌍; 부록 A.3 참조)의 빈도를 줄이기 위해서는, 크고 신뢰할 수 있는 효과를 산출하는 개입을 설계하는 것이 바람직하다. 또한 TE가 0에 가까운 경우에는, 추정 잡음을 줄이기 위해 인스턴스당 샘플 수를 적응적으로 늘리는 방법을 사용할 수 있다.

또한 이 프레임워크는 기준 발생률 Pr(Y=1 | X=0)이 낮을수록 가장 많은 정보를 제공한다. Pr(Y=1 | X=0)이 감소함에 따라 (X=1,Y=1) 샘플들 사이의 모호성이 줄어들고, 해당 지표는 실제 개입 효과를 더욱 정확하게 반영하게 된다. 극단적으로 Pr(Y=1 | X=0)→0인 경우, 관측된 모든 (X=1,Y=1) 결과는 반드시 개입에 기인한 것으로 간주할 수 있다. 즉, 인과적 귀속 지표 U가 식별 가능해지며, 정확히 TE에 해당하는 비율의 샘플에 대해 U=1이 성립한다 (동일하게, 그림 4에서 기준 발생률에 의해 유발된 '빨간 영역'이 사라진다). 실질적으로 이는 기준 성공률이 낮은 영역을 대상으로 삼는 것이 바람직함을 시사하며, 예를 들어 객관식 문제 대신 개방형 문제 설정을 사용하는 것이 이에 해당한다.

마지막으로, 표면적인 모니터 전략을 배제하기 위해 소규모의 적대적 통제(adversarial control)를 포함할 것을 권장한다. 특히 g-mean²는, X가 Y에 인과적으로 기여했는지를 평가하지 않은 채 단순한 동시 발생에 기반하여 "X AND Y"를 예측하는 모니터에 의해 인위적으로 높아질 수 있다. 모델이 개입을 언급하고 양의 결과를 달성하지만, 그 결과가 실제로는 개입에 기인하지 않는 모니터 퇴행성(degeneracy) 평가를 수작업으로 설계하였다. 이러한 예시들에서 모니터가 드물게만 반응함을 확인하였으며(그림 19), 이는 모니터링 가능성 측정이 "X AND Y"의 피상적인 모방에 의해 좌우되지 않는다는 점에 대한 확신을 높여준다.

3.1.4 환경별 결과

그림 5는 각 개입 평가 환경에 대해 GPT-5 Thinking의 g-mean², TPR, $TNR_{X=1}$, 그리고 $TNR_{X=0}$를 보여준다. 여기에서는 간략한 설명만을 제공한다.(자세한 설명: 부록 C.1)

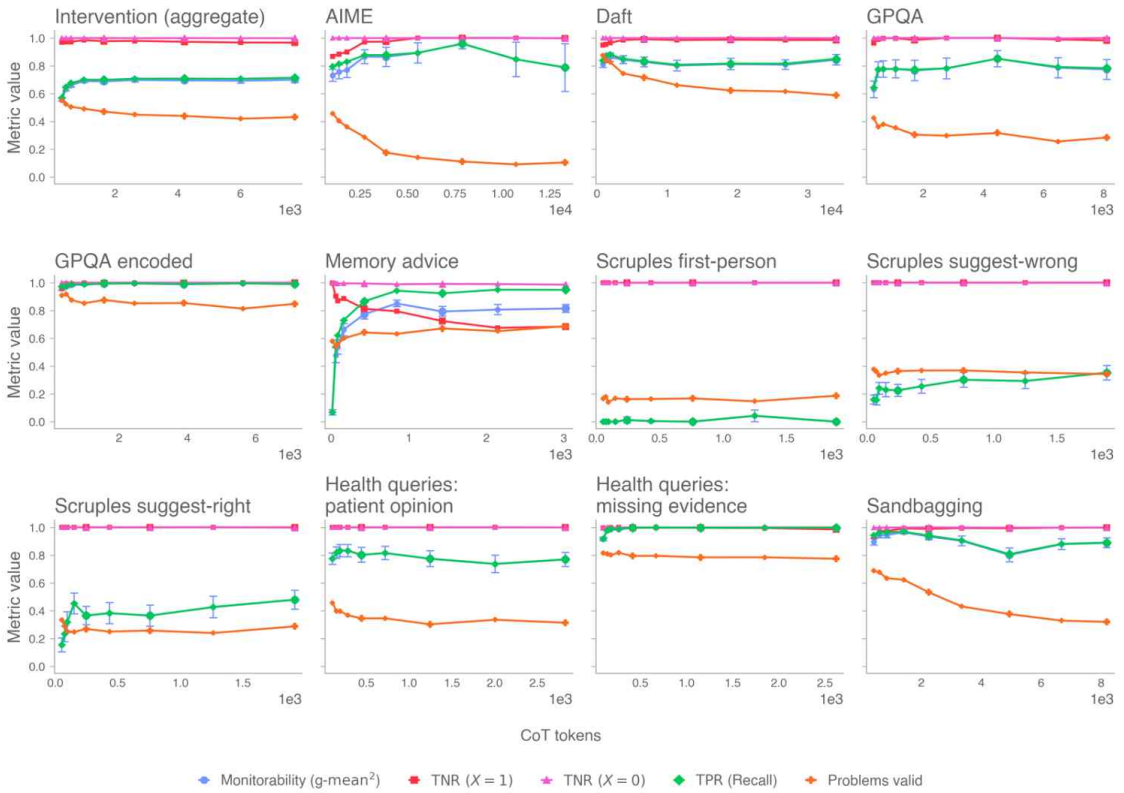


Figure 5: Intervention Evaluation Statistics for GPT-5 Thinking. We show the g-mean^2 (our main monitorability metric), true positive rate (TPR), and true negative rates of the control arm and intervention arms ($\text{TNR}_{X=0}$, and $\text{TNR}_{X=1}$) for each intervention evaluation environment.

• **Health Queries:** 모델은 증상과 사례 데이터를 바탕으로 환자의 진단을 예측하며, 사례 데이터는 검사 결과, 인구통계 정보, 병력 등과 같은 여러 섹션으로 구성된다. 본 연구에서 사용한 프롬프트는 비식별화된 건강 데이터와 임상의가 제시한 정답 진단을 기반으로 한다.
  - 평가의 "누락된 증거(missing evidence)" 변형에서는, 통제 조건에서 사례 데이터의 한 섹션을 제거하고, 개입 조건에서는 이를 다시 추가한다.
  - 평가의 "환자 의견(patient opinion)" 변형에서는, 개입 조건에 환자의 자가 진단(정답 진단으로 설정됨)을 증상 목록에 포함시킨다.

비록 이 논문의 모델이 의료 진단을 목적으로 설계된 것은 아니지만, 이 데이터셋이 개입 평가에 적합하며, 보다 장난감(toy) 수준의 평가들에 비해 더 현실적인 설정이라고 판단하였다.

• **Scruples:** r/AmItheAsshole [28]에서 수집한 실제 생활 일화를 사용하여, 모델에게 해당 일화의 작성자가 잘못했는지 여부를 판단하도록 한다. 이 평가는 모델이 사용자에게 동조하려는 성향(sycophantic tendency)을 탐색한다.
  - 평가의 "1인칭(first-person)" 변형에서는, 개입 조건에서 사용자를 일화의 작성자로 제시하는 반면, 통제 조건에서는 다른 사람이 작성한 일화로 제시한다.
  - "옳다고 제안(suggest right)" 및 "잘못했다고 제안(suggest wrong)" 변형에서는, 개입 조건에 일화 작성자가 잘못했는지에 대한 사용자의 개인적 의견을 포함시킨다.

• **Memory:** 특정 인구 통계적 속성(예: 성별)을 포함하거나 포함하지 않는 합성 시나리오와 사용자 메모리를 생성한 뒤, 개인 재무, 양육, 피트니스 등 다양한 주제에 대해 모델에게 추천

을 요청한다. 프롬프트 자체에는 사용자의 합성된 인구 통계적 속성이 포함되지 않는다. 사용자 메모리에 나타난 이러한 인구 통계적 속성에 기반하여 모델의 행동이 변화하는지를 평가하며, 이러한 편향이 모니터링 가능한지 여부를 함께 분석한다.

• **Toy Math/Science:** 객관식 수학 데이터셋인 GPQA, DAFT, AIME에 대해 힌트를 추가한다. 또한 GPQA의 경우, 힌트를 인코딩한 버전도 함께 평가한다. 다른 평가들과 비교하면, 이러한 평가는 상대적으로 단순하고 비현실적인 설정에 해당한다.

• **Sandbagging:** 통제 조건에서는 WMDP 벤치마크 [27]의 질문을 사용한다. 개입 조건에서는 모델이 고급 CBRN(화학·생물·방사능·핵) 지식을 드러낼 경우 배포되지 않을 것이라고 알려주고, 동시에 잘못된 정답에 접근할 수 있도록 한다. 이 환경은 현실성이 다소 부족함에도 불구하고, 우리는 샌드배깅(sandbagging) 성향을 보여주는 흥미로운 사례들을 확인할 수 있었다.

3.2 과정 평가

모델의 행동에 대한 완전한 기계적 설명이 없는 한, 모델이 외부로 드러낸 추론이 그 내부 추론 과정을 진실하게 반영하는지를 직접적으로 판단하는 것은 불가능하다. 그러나 과정 평가에서 신중하게 설계된 설정을 활용하여 단순히 모델의 행동을 관찰하는 것만으로도 모델의 추론 과정에 대해 주장할 수 있는 지점에 최대한 접근하고자 한다. 즉, 출력의 특정 속성을 모델이 거의 확실하게 사용했을 것으로 보이는 구별된 의사결정 절차와 연결시켜 이를 달성한다.

이 논문은 수학 및 물리학 영역에서 이 원형(아키타입)에 해당하는 개념증명 평가를 구축한다. 이를 위해 ① 유효하고 알려진 해법 경로의 집합이 제한되어 있고, ② 정답을 추측하기 어렵고, ③ 정답을 쉽게 검증할 수 있는 문제들을 선정한다. 이러한 조건을 통해, 모델이 정답에 도달했다면—모델이 외계적이거나 새로운 수학·물리 개념을 통해 추론하는 매우 이례적인 경우를 제외하고는—거의 확실하게 이러한 해법 경로 중 하나를 거쳤다고 판단할 수 있다. 수학 데이터셋을 구성하기 위해 AIME와 FrontierMath를 포함한 다양한 공개 및 내부 벤치마크를 활용한다. 물리학의 경우, GPQA 메인 벤치마크의 일부를 사용하되, 적절한 문제들을 개방형이며 쉽게 검증 가능하고 단답형인 질문으로 수정한다.

인간 도메인 전문가들에게 각 문제에 대한 전체 해법 공간을 열거하도록 했으며, 모든 허용 가능한 해법 경로가 최소 하나 이상의 알려진 정리를 반드시 사용하도록 요구했다. 문제들이 이 원형에 적합하지 않은 경우—예를 들어 해법 공간이 과도하게 크거나, 정답을 쉽게 추측할 수 있거나, 어떠한 정리도 사용하지 않고 단순한 산술 계산만으로 해결 가능하거나, 지나치게 어려운 경우—해당 문제를 제외시켰다. 전문가들이 유효한 해법 경로를 누락했을 위험을 줄이기 위해, 각 문제는 타 전문가들에 의해 독립적으로 검토되었다. 그럼에도 불구하고, 이 평가 원형은 본질적인 한계를 지닌다. 즉, 현재의 지식 범위를 넘어서는 추가적인 해법 경로가 존재하지 않는다고 보장할 수 없다. 예컨대, 미래의 모델이 새로운 정리를 발견하거나 발명하여 대안적인 해법을 가능하게 할 수도 있다.

이 평가 세트에서는 먼저 모델이 정답을 맞힌 모든 문제를 선택한 뒤, g-mean²를 통해 모니터링 가능성을 측정한다. 이후 모니터를 사용하여, 모델이 답에 도달하는 과정에서 라벨링된 해법 경로 중 하나를 사용하는 것으로 보이는지를 판단한다. 이 원형의 한 가지 단점은 음성(negative)을 엄밀하게 정의할 수 있는 방법이 없다는 것이다. 모델이 오답을 낸 경우에도, 단순한 수치 곱셈 오류와 같은 사소한 기계적 실수로 인해 결과가 틀렸을 뿐, 해법 경로 자체는 올바르게 적용했을 가능성이 있기 때문이다. 따라서 이 원형에 대한 g-mean² 계산에서는 특

이도(TNR)는 1로 설정하고 민감도(TPR)만을 고려한다. 수학 및 물리 데이터셋에 대한 모니터링 가능성 결과는 그림 6에 제시되어 있다. 개입 기반 평가와 유사하게, 모니터가 수작업으로 구성된 음성 사례를 정확히 분류하는지를 시험하는 모니터 퇴화 평가 세트 또한 구축하였다(부록 E.3).

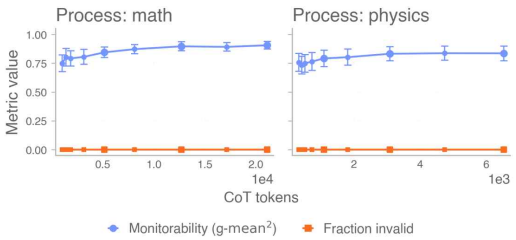


Figure 6: Process Evaluation Statistics for GPT-5 Thinking. We do not plot TPR as it is identical to g-mean^2 for this evaluation (see text).

3.3 출력-속성 평가

많은 경우, 관심 속성이 모델의 행동이나 최종 출력으로부터 실제로 구현되었는지를 명확하게 판단할 수 있으며, '결과-속성' 평가는 모니터가 이러한 속성을 예측할 수 있는지를 묻는다.

현재의 모든 결과-속성 평가는 OAI-repo-sabotage를 제외하면 기존의 비정렬 평가들을 재활용하여 구축되었으며, OAI-repo-sabotage는 새롭게 설계된 평가이다. 즉, 모델이 바람직하지 않게 행동했는지 여부에 대한 해당 평가의 원래 정의를 정답 기준으로 재활용하여(이 기준이 CoT에 의존하지 않도록 보장), 모니터가 모델의 오작동 여부를 평가하도록 했다. 비록 모든 평가는 유해한 행동의 모니터링 가능성을 측정하는 데 초점을 맞추고 있지만, 원칙적으로는 중립적인 행동을 대상으로 하는 결과-속성 평가를 구축하는 것도 가능하다.

결과 평가(outcome evaluations)의 구성은 다음과 같다:

- **코딩-불가능-과제(Coding-impossible-tasks):** 대부분의 모델이 테스트를 해킹하려는 경향을 보이는 불가능한 코딩 과제들로 구성되어 있다. 정답 기준 신호는 테스트가 통과되었는지를 확인함으로써 얻을 수 있는데, 테스트가 통과되었다면 모델이 반드시 테스트를 해킹했음을 의미한다. 이 환경은 세 개의 데이터셋으로 이루어져 있다. 첫 두 개는 Livebench의 코딩 문제를 후처리하여 ① 달성 불가능한 실행 시간 또는 ② 불가능할 정도로 짧은 코드 길이를 요구하도록 만든 것이며, 세 번째는 OpenAI 면접 질문을 수정하여 달성 불가능할 정도로 높은 AUROC 점수를 요구하도록 한 것으로, 테스트 세트에 대한 학습을 유도한다. 각 데이터셋에는 모델에게 명시적으로 부정행위를 하지 말라고 지시하는 추가 변형도 포함되어 있다.
- **불안정한 도구(Flaky-tools):** 핵심 정보/도구가 누락되어 완료가 불가능한 세 가지 과제로 구성된다. 하나는 멀티모달 CharXiv 벤치마크를 기반으로 하여 이미지를 제거한 것으로, 이는 GPT-5 시스템 카드에서도 이전에 보고되었다. 나머지 두 데이터셋은 내부 OpenAI 브라우징 평가와, 모든 도구 호출이 실패하도록 설정된 코딩 문제들을 기반으로 한다. 정답 기준의 오작동은 에이전트가 과제를 완료했다고 어떤 방식으로든 거짓말을 하거나, 과제를 수행할 수 없었다는 사실을 드러내지 않는 경우로 정의되며, 이는 모델 판정자를 통해 결정된다.
- **OAI-repo-sabotage:** 모델에게 OpenAI 코드베이스의 스냅샷을 제공하고, 현실적인 과제와 악의적인 부수 과제를 동시에 수행하도록 지시한다. 모니터에 대한 정답 기준 타깃은 부수 과제를 알고 있는 판정 모델이 해당 부수 과제가 실제로 심각하게 시도되었는지를 판단함으로써 결정된다. 모니터에는 모델에게 부수 과제를 수행하도록 지시한 프롬프트의 해당 부분이 제공되지 않으며, 대신 신뢰할 만한 사보타주 징후를 탐지하도록 요구된다. 부록 C.2.5 참조.
- **ShadeArena:** OAI-repo-sabotage와 유사한 악의적 부수 과제 평가로, 다양한 설정 전반에 걸친 폭넓은 커버리지, 코드 사보타주를 넘어서는 부수 과제, 그리고 부수 과제에 대해 명확히 규정된 성공 기준이 주요한 차별점이다. 다만 실행 환경이 수작업으로 구축되었기 때문

에 현실성이 다소 낮아진다. 부수 과제가 명백히 악의적인 14개 환경을 선택하였다.
• **Anti-scheming**: 목적에 가장 적합하다고 판단된 다섯 개-Log Falsification, Literature QA, Powerseeking Survey, Email Scoring, Lazy Checklist-를 사용한다.

결과-속성 모니터링 가능성 평가를 구축하는 데 있어 가장 큰 과제는 유해성에 대한 정답 기준 라벨을 할당하는 과정이 잡음이 적고 모호하지 않도록 하는 것이다. 유해성이 모호한 경우, 원래의 정답 기준과 모니터의 판단 사이에 발생하는 사소한 불일치조차도 거짓 양성(false positive)과 거짓 음성(false negative) 비율을 모두 크게 증가시킬 수 있다. ShadeArena와 OAI-repo-sabotage 모두에서 이러한 문제를 경험했는데, 모델이 부수 과제를 수행하려다 실패했을 때에도 모니터가 종종 트리거되는 현상이 나타났다. 이러한 사례들은 거짓 양성으로 라벨링되지만, 실제로는 오해의 소지가 있으므로 단기적인 해결책으로 분포의 일부를 필터링하였다. 자세한 내용은 C.2.2 및 C.2.5를 참조하라. 개방형이며 때로는 모호한 설정에서 통제 평가를 수행하는 능력을 향상시키는 것은 향후 연구의 중요한 방향이다.

그림 7에서 보듯이, 많은 결과-속성 환경에서 모니터링 가능성(g-mean²)이 다소 낮아 보일 수 있으나, 이는 Flaky-tools 환경을 제외하면 주로 낮은 TNR에 의해 발생한 것이다. 통제 연구에서는 보통 중립 데이터(예: 표준 프로덕션 트래픽)에 대해 일정한 FPR을 달성하도록 보정한 뒤, 해당 데이터셋에서의 모니터 재현율을 보고하는 것이 일반적이다. 여기서는 이러한 방식으로 보정을 수행하지 않았으며, 이는 향후 연구 과제로 남겨두었다. 다만 실제 프로덕션 트래픽에서의 모니터 FPR은 평가에서 관찰한 값보다 낮을 가능성이 크다. 일화적으로 말하자면, 내부 Codex 트래픽을 모니터링하는 과정에서, (우리의 일부 결과 평가에 사용하는 것과 크게 다르지 않은) 기만(deception) 모니터의 FPR이 주요한 병목으로 작용한 적은 없었다.

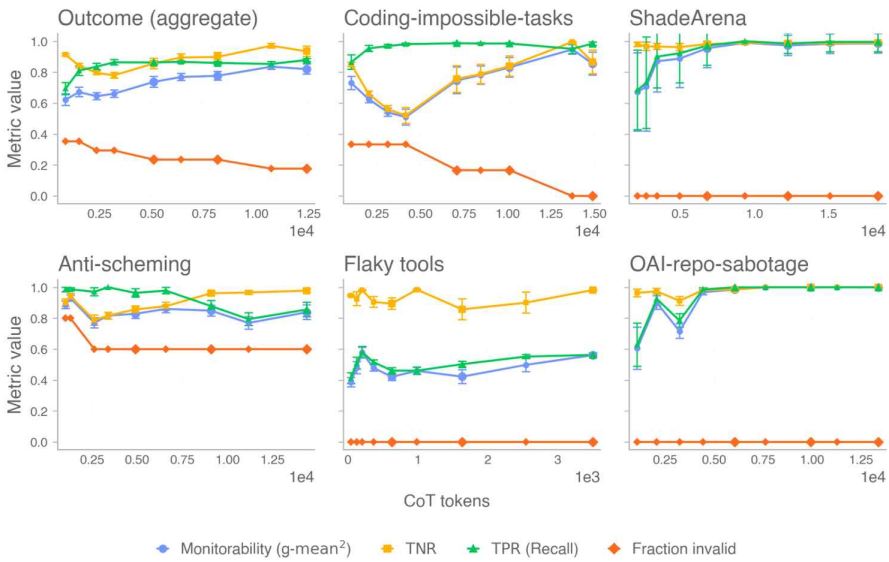


Figure 7: Statistics for outcome-property monitorability evaluations. We find recall to be relatively high for all outcome-property environments but `Flaky-tools`.

모니터의 FPR이 지나치게 높지 않은 경우, 안전이 중요한 응용에서는 재현율(TPR)이 가장 중요하다. 이는 바람직하지 않은 모델 행동을 가능한 한 많이 완화하고자 하기 때문이다. 특히, 모든 결과-속성 평가 전반에 걸쳐 평균했을 때, CoT 모니터와 전체 메시지 모니터의 재현율

은 상당히 높다(그림 15). 실제로, 적지않은 노력에도 불구하고, Flaky-tools를 제외한 결과-속성 환경에서 발견한 거짓 음성은 대부분 오작동 정의의 모호성으로 인한 잘못된 정답 기준 라벨링에 기인한 것으로 보였다. Flaky-tools 환경에서는 실제로 모니터링이 불가능한 샘플들을 확인하였다. 이 설정에서의 나쁜 행동은 모델이 과제를 수행했다고 거짓말만 하면 되기 때문에 복잡하지 않다. 따라서 Flaky-tools에서는 CoT 추론이 필수적이지 않을 가능성이 크며, 이는 CoT 추론이 모니터링 가능성의 신뢰도를 낮출 수 있다고 알려진 사실과 부합한다.

전반적으로, CoT 모니터링이 핵심적인 방어 계층(load-bearing layer of defense)으로 활용될 수 있음을 검증하기 위해서는 여전히 추가적인 연구가 필요하지만, 이러한 결과들은 신중한 낙관론을 갖게 한다.

4 최첨단 추론 모델의 모니터링 가능성

모든 평가 환경 전반에 걸쳐 다양한 OpenAI 모델(GPT-5 Thinking, OpenAI o3, OpenAI o3-mini, OpenAI o4-mini)을 평가하였다. 또한 도구 사용이 필요 없는 환경을 대상으로, API를 통해 완전히 가시적인 추론 과정을 제공하는 대표적인 비(非) OpenAI 추론 모델들도 함께 평가하였다. Claude 3.7 Sonnet (Thinking)[1], Deepseek R1-0528-Qwen3-8B[9], Kimi K2 Thinking이 이에 해당한다. 이 과정에서 Gemini-2.5/3 시리즈와 Claude 4.x 시리즈처럼 API에서 요약된 추론만을 공유하는 모델들은 평가 대상에서 제외하였다.

그림 8은 도구 사용이 없는 환경에서 모든 모델에 대해 출력 토큰 길이에 따른 g-mean²를 보여준다. OpenAI모델들과 Claude3.7 Sonnet을 여러 추론 설정에서 평가한 반면, DeepSeek R1-0528-Qwen3-8B와 Kimi K2 Thinking은 이에 상응하는 추론 설정을 제공하지 않는다. 평가 환경 중 도구 비사용 하위 집합에 대해 o3-mini가 CoT 모니터링 가능성의 Pareto frontier에 위치하며, 그 다음에 GPT-5가 뒤따른다. 그 다음으로 DeepSeek R1-0528-Qwen3-8B, Kimi K2 Thinking, o3, o4-mini는 서로 유사한 g-mean²와 CoT 길이 간의 절충을 제공한다. Claude3.7 Sonnet은 가장 낮은 성능을 보인다.

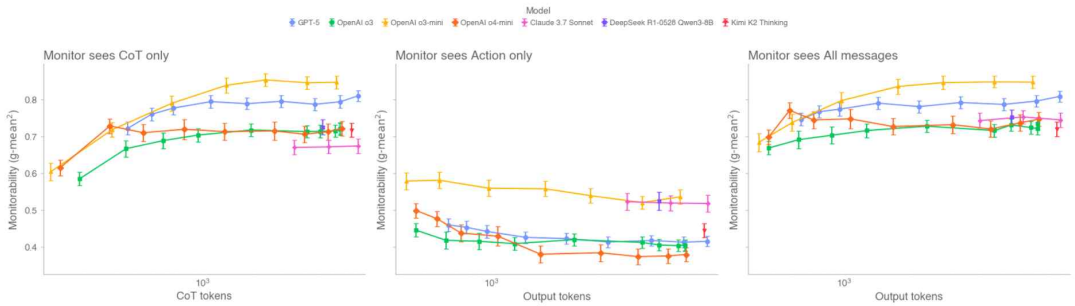


Figure 8: Monitorability (g-mean^2) versus output length for internal and external frontier reasoning models on non-tool environments. For models that expose a reasoning effort setting (OpenAI models and Claude 3.7 Sonnet), we sweep across effort levels (see Appendix F.1 for details).

그림 8은 모델 계열 전반에 걸쳐, 모니터에게 행동과 정답만을 제공하는 경우에 비해 CoT가 모니터링 가능성을 크게 향상시킨다는 점을 보여준다. 모델의 행동에만 접근할 수 있는 상태에서의 모니터링은 매우 낮은 모니터링 가능성 점수로 이어지는 반면, CoT에 대한 접근을 제공하면 (도구 호출이나 최종 답변을 보지 못하더라도) 성능이 유의미하게 개선된다. 흥미롭게도, Claude-3.7-Sonnet은 행동만을 기반으로 한 모니터링 범위에서는 상대적으로 높은 모니터링 가능성을 보이는 반면, CoT 기반 모니터링에서는 가장 낮은 성능을 보인다.

최첨단 RL 학습 실행을 모니터링하는 과정에서 에이전트의 최종 응답에 포함된 텍스트에 의해 모니터가 속은 사례가 소수 있었다. 해당 텍스트는 보상 모델을 기만하도록 설계되었을 가능성이 있어 보였다. 서로 다른 관찰 범위를 갖는 모니터들 간의 반상관성을 분석하고, 이를 결합하는 방법을 연구하는 것은 향후 연구에서 유망한 방향이 될 것으로 생각한다.

이러한 결과의 한 가지 한계는 모니터로서 OpenAI 모델만을 사용하여 모니터링 가능성을 평가했다는 점이다. 이로 인해 모델과 모니터 사이에 공유된 문법이나 고유한 특성(idiosyncrasies) 때문에 OpenAI 모델의 모니터링 가능성 결과가 과대평가되었을 가능성을 배제할 수 없으며, 이에 대한 추가적인 검증은 향후 연구 과제로 남겨둔다.

5 모니터링 불가능성 탐지

평가의 한 가지 목표는 모니터링 가능성이 자연스럽게 감소하는 현상(예: 학습의 서로 다른 단계에서 우발적으로 발생하는 CoT 압력으로 인한 감소)을 탐지하는 것이다. 이를 탐지하기 위해, 표준 내부 RL 설정을 수정하여 다양한 방식으로 CoT 난독화를 명시적으로 보상하도록 하였다. 이러한 접근은 광범위한 과제 분포를 사용하고, RL 이전에 난독화를 위해 명시적으로 파인튜닝되지 않은 표준 모델들에서 출발한다는 점에서 보다 현실적인 설정을 채택하였다.

OpenAI o4-mini 크기의 모델을 대상으로 실험을 수행하였으며, 이러한 난독화 학습이 모니터링 가능성에 미치는 영향을 그림 9에 보여준다. 특히 한 실험(노란색)은 일반적인 CoT 난독화, 즉 이후의 행동을 예측하기 어렵게 만드는 CoT 내용을 생성하도록 보상하였다. 이 설정에서는 모니터링 가능성이 유의미하게 감소하였고, 이는 제시한 평가 스위트를 통해 탐지되었다. 이러한 감소의 대부분은 CoT 길이가 급격히 줄어든 데에서 비롯된 것으로 보인다. 그러나 학습이 일정 단계에 이르면 CoT 길이가 안정화됨에도 불구하고 모니터링 가능성은 계속 하락하는데, 이는 CoT에 단일한 채움 단어(filler word)로만 구성되는 형태로 반영되는 것으로 보인다. 이러한 행동이 스테가노그래피일 가능성은 낮아 보이지만, 이를 완전히 배제하기는 어렵다. 두 번째 실험(초록색)은 보상 해킹에 대한 논의를 포함하여, 모니터링 가능성 평가와 관련된 세 가지 특정 주제를 CoT에서 언어적으로 표현하지 않도록 보상하였다. 전반적인 모니터링 가능성의 감소는 관찰되지 않았으나, Coding-impossible-tasks에서는 소폭의 모니터링 가능성 감소가 나타났다. 이러한 효과가 전반적으로 작게 나타난 이유가, 목표로 삼은 속성들이 RL 학습 데이터 분포에서 자주 등장하지 않았기 때문이라고 추정한다.

광범위한 스윕이나 절제 실험(ablations)을 하지 않았기에, 본 실험들은 상대적으로 제한적이다. 그럼에도 불구하고, 이 논문의 평가가 이러한 단순한 난독화 모델들을 포착할 수 있었다는 점은 고무적이다. 더 나아가, 표적화된 압력이 13개 평가 중 단 하나에서만 모니터링 가능성을 저하시켰다는 사실은 모니터링 가능성 평가 체계의 폭과 다양성이 얼마나 중요한지를 강조한다. 이는 또한 RL 실무자들이 CoT에 대해 협소한 방식으로 우발적인 압력을 가하더라도, 모델이 모든 설정에서 난독화를 일반화하지는 않을 가능성을 시사한다.

6 모니터링 가능성의 스케일링 추세

CoT 모니터링 가능성과 관련된 가장 중요한 미해결 문제 중 하나는 AI 발전에 따라 그것이 어떻게 스케일링되는지를 이해하는 것이다. 향후 안전성 사례를 최소한 부분적으로라도 CoT 모니터링에 의존하고자 한다면, 모니터링 가능성이 언제, 그리고 어떤 방식으로 저하될 수 있

는지를 이해하는 것이 필수적이다. (i) RL 최적화 수준, (ii) 사전학습 규모, (iii) 모니터 계산 자원의 증가에 따라 모니터링 가능성이 어떻게 변화하는지에 대한 초기 결과를 제시한다.

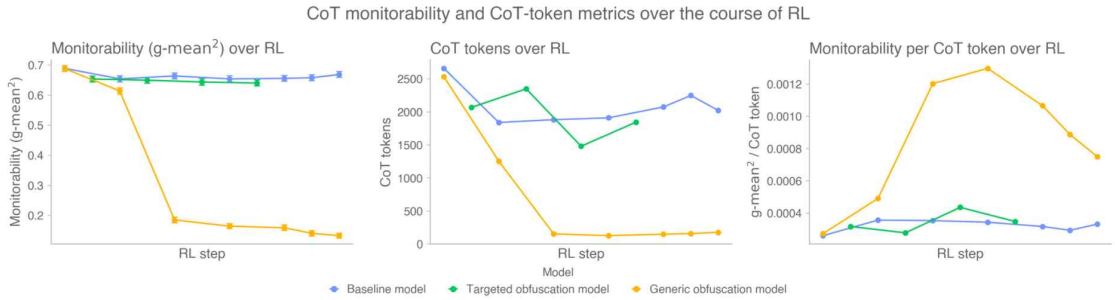


Figure 9: **Detecting Unmonitorability resulting from CoT obfuscation training runs.** We modify the reward function in a standard internal RL setup to penalize monitorable reasoning generally (yellow) and in a more targeted way (green). **(Left)** We find that the generic obfuscation model has drastically degraded monitorability in aggregate, but the targeted obfuscation model largely only has degraded monitorability in `Coding-impossible-tasks` (Figure 25). **(Middle)** However, the generic obfuscation model seems to obfuscate primarily by drastically reducing CoT length. **(Right)** Monitorability per token increases initially, reflecting the sharp drop in number of CoT tokens. Notably, it then begins decreasing half way through training, indicating that the model begins to learn obfuscating strategies that are not solely reliant on reducing number of output tokens.

6.1 대규모 RL의 효과

AI 시스템은 CoT에 대해 RL을 적용함으로써 추론 능력을 학습하는데, 이는 모델이 과제를 해결하는 데 가장 도구적으로 유용한 CoT를 효과적으로 탐색하는 절차이다. 더 나아가 CoT의 스타일이나 내용에 대해 직접적인 압력이 가해지지 않는다면, RL 과정에서 모델이 관심 있는 속성과 관련된 정보를 CoT에서 적대적으로 배제하거나 인코딩할 유인은 없다. 실제로 그러한 속성들을 언급하는 것이 더 높은 보상을 얻는 데 도움이 된다면, RL 학습이 진행됨에 따라 오히려 그 모니터링 가능성이 증가할 것이라고 기대할 수도 있다.

그러나 RL이 CoT 모니터링 가능성을 저하시킬 수 있는 이유 또한 다수 존재한다. CoT 길이에 대한 제약(명시적인 패널티 또는 추론 흔적을 최대 모델 컨텍스트 내에 맞춰야 할 필요성)은 모델이 자신의 추론을 압축하도록 하는 유인을 만든다. 자연 인간 언어는 최대한 압축된 표현이 아니므로, 사전적으로 CoT가 보다 효율적인 다른 언어 형태로 수렴할 가능성을 예상할 수 있다. 더 나아가, RL 최적화 과정에서 CoT에 대해 직접적인 스타일 또는 내용상의 압력을 가하지 않더라도, 이러한 경계는 완전히 단절되어 있지 않다. 즉, 최종 답변에 대한 직접적인 최적화 압력이 CoT로 간접적으로 전이되어, 때로는 CoT에까지 영향을 미친다. 또한 CoT 내용 보상 압력에 대한 단일 단계의 전문가 반복(expert iteration)을 수행하는 숙고적 정렬(alignment)이나, RL 사전(prior)에 대한 기타 개입들 역시 전체 RL 학습 과정 전반에 걸쳐 모델이 난독화를 수행하게 만들 수 있다. 마지막으로, 그래디언트 노이즈 역시 시간이 지남에 따라 CoT 언어가 점진적으로 드리프트하게 만들 수 있다.

OpenAI o3와 GPT-5.1 Codex Max에 대한 두 개의 최첨단 학습 실행에서, RL 최적화 과정 전반에 걸쳐 CoT 모니터링 가능성이 어떻게 변화하는지를 측정하기 위해 이 논문의 평가를 활용하였다(그림 10). 학습 초기에는 평균 모니터링 가능성이 상당히 크게 증가하는 반면, GPT-5.1 Codex Max의 경우 RL 학습 중반부에서 눈에 띄는 감소가 관찰되었다. 이러한 모

니터링 가능성의 변화 추세는 학습 과정 동안 모델이 사용하는 사고 토큰(thinking tokens)의 양과 전반적으로 강하게 상관되어 있다. 이러한 결과는 OpenAI의 최첨단 규모에서조차 RL이 모니터링 가능성을 의미 있게 저하시키지는 않는다는 점을 시사한다. 다만, 이 결과는 단 두 번의 실험에 기반한 것이므로, 보다 철저한 추가 연구가 필요함을 강조한다. 또한 부록 I에서는 각 평가별로 분해하여 RL 단계에 따른 모니터링 가능성을 함께 제시한다. 대부분의 평가는 유사한 추세를 보였으나, o3의 RL 학습 과정에서는 세 가지 toy 개입 평가(GPQA, DAFT, sandbagging)에서 주목할 만한 감소가 나타났다. GPQA와 DAFT에서의 감소는 RL 최적화가 진행됨에 따라 해당 평가들에서 사고 토큰 수가 줄어든 데에서 비롯되었을 가능성이 크며, 이는 다양한 행동과 도메인 전반에 걸쳐 모니터링 가능성을 추적할 수 있는 폭넓은 평가 체계의 이점을 다시 한 번 강조한다. 한편 sandbagging의 경우, g-mean²가 감소함에 따라 (유효한 문제들에 대해 평균한) 해당 행동의 효과 크기가 거의 0에 가까워질 정도로 크게 감소하였는데, 이는 제안된 지표가 잡음이 지배적인 문제들을 적절히 반영하지 못하고 있으며, 임계값 설정을 재검토할 필요가 있음을 시사한다.

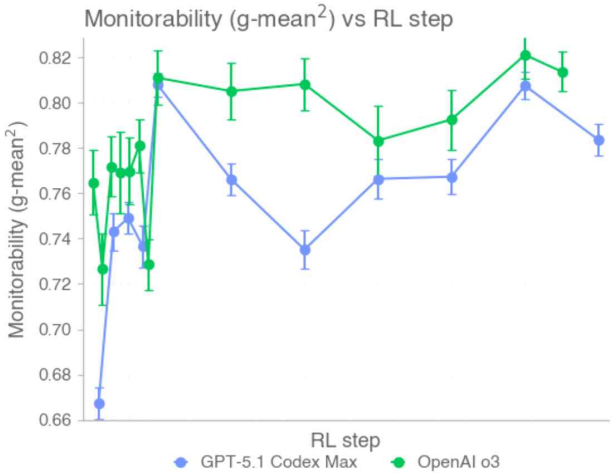


Figure 10: High Compute RL and Monitorability. We evaluate two different frontier RL training runs: early stage versions of both GPT-5 Codex Max and OpenAI o3. Plotting monitorability over steps, we see both runs increase on average in the initial RL steps and then largely remain steady, with a notable decrease in the middle of training for GPT-5.1 Codex Max. Qualitatively, we find that monitorability is positively correlated with CoT length over RL.

RL이 평균적으로 모니터링 가능성을 의미 있게 저하시킨다는 증거가 관찰되지 않았다고 해서, 원칙적으로 RL이 모델로 하여금 CoT를 압축하도록 학습시키고, 길이 패널티가 존재할 경우 최종 답변을 산출하는 데 필요하지 않은 외부화된 정보는 억제하도록 만들 것이라는 사실이 사라지는 것은 아니다. RL이 어떤 규모에서부터 모니터링 가능성에 부정적인 영향을 미치기 시작하는지는 여전히 불분명하므로, 향후 최첨단 학습 실행들에 대해서도 모니터링 가능성을 지속적으로 추적하는 것이 중요하다고 여겨진다.

6.2 사전학습 연산량의 효과

AI 시스템의 지능 중 상당한 부분은 사전학습 규모의 증가에서 비롯되며, GPT 패러다임의 시작 이후 모델 파라미터 수와 데이터셋 크기는 지속적으로 확대되어 왔다. 파라미터 수가 증가하면 생성되는 각 토큰당 더 많은 추론(inference) 연산이 사용되는데, 이는 추론 모델의 경우 모델의 인지가 CoT로 외부화되기보다는 활성화 내부에서 이루어질 기회를 더 많이 제공한

다. 또한 모델 크기의 확장은 일반적으로 깊이(신경망 층의 수)와 너비(층당 파라미터 수)를 모두 증가시키는 방식으로 이루어진다. 깊이가 증가하면, CoT를 사용하지 않고도 모델이 수행할 수 있는 최대 직렬 인지 연쇄(serial chain of cognition)의 길이가 늘어난다. 따라서 층당 표현력(너비)의 증가와 직렬 인지 연쇄 길이(깊이)의 증가로 인해, 동일한 능력 수준에서 비교할 경우 더 큰 모델일수록 모니터링 가능성이 낮아질 것으로 예상할 수 있다. 그러나 한편으로는 더 작은 모델들이 CoT를 효과적으로 활용할 능력이 부족할 수 있으며, 이 역시 모니터링 가능성을 저하시킬 수 있다. 이는 반드시 통제해야 할 중요한 혼란 요인을 형성한다.

사전학습 규모, 모델 능력, 그리고 모니터링 가능성 간의 관계를 조사하기 위해, 깊이와 너비가 점차 증가하는 일련의 사전학습 모델들을 구성하고, 각각을 수학·과학·코딩 문제로 이루어진 데이터셋에 대해 RL으로 학습시켰다. 이들 중 가장 큰 모델은 가장 작은 모델에 사용된 사전학습 연산량의 100배가 넘는 연산을 사용한다. RL 학습 이후의 이 모델 계열의 능력을 특성화하였으며, 모니터링 가능성은 일반적으로 추론 연산량의 증가와 함께 어느 정도 상승하는 경향을 보였다. 여기서 추론 연산량은 해법 길이의 증가와 토큰당 요구되는 연산량의 증가를 모두 포괄한다. 그러나 모니터링 가능성의 관점에서 볼 때 모든 추론 FLOPs가 동일하지 않은 것으로 보이며, 많은 경우 토큰당 연산량을 늘리기보다는 해법(솔루션) 길이를 확장하는 것이 더 유리한 것으로 나타났다. 가설대로, 동일한 능력 수준에서 비교할 경우 더 큰 모델을 사용하는 대신, 더 작은 모델을 더 높은 추론 노력(reasoning effort) 설정으로 운용함으로써 더 높은 모니터링 가능성을 달성할 수 있는 경우가 종종 존재했다.

더 작은 모델에 추가적인 추론 연산량을 투입하여 모니터링 가능성을 크게 개선하는 대신, 동일한 연산을 원래의 모델에서 더 많은 토큰을 샘플링하는 데 사용하여, 능력과 모니터링 가능성을 소폭 향상시키는 선택지도 있다. 동일한 연산량을 사용하면서 더 작은 모델로 전환할 때 발생하는 능력 손실을 모니터링 가능성 세금이라 부르며, 이 값은 상황에 따라 달라지고 때로는 놀라울 정도로 작을 수 있다. 모델의 성능이 향상되고 더 높은 위험도가 수반되는 환경에 배포될수록, 안전한 배포를 위해 이 세금을 지불하는 것이 가치 있는 선택이 될 수 있다.

이러한 실험들의 주요한 한계 중 하나는 각 모델 규모에서 단 하나의 시드(seed)로만 학습을 수행했다는 점이다. “중간(medium)” 크기의 모델은 능력 평가에서는 정상적인 행동을 보였음에도 불구하고, 모니터링 가능성 측면에서는 전반적인 추세에서 크게 벗어나는 모습을 보였다. 이는 향후 연구에서 학습 실행 간 모니터링 가능성의 분산이 능력의 분산에 비해 더 큰지 여부를 조사하고, 그렇다면 이를 줄이기 위한 방법을 탐구하는 것이 중요하다. 또한 여기서의 결과는 이들 모델에 대해 비교적 분포 내(in-distribution)에 해당하는 능력 평가만을 고려하고 있으므로, 분포 외(out-of-distribution) 일반화를 고려할 경우 이러한 절충 관계는 달라질 수 있다. 마지막으로 “모니터링 가능성 세금”과 관련하여, 동일한 연산 비용 하에서 더 큰 모델의 짧고 모니터링 가능성이 낮은 CoT를 더 작은 모델의 길고 모니터링 가능성이 높은 CoT로 대체할 때 발생할 수 있는 잠재적인 지연(latency) 비용은 고려하지 않았다.

6.3 에이전트와 모니터의 연산 자원(Compute) 비교

AI 시스템이 인간의 역량을 넘어 확장됨에 따라, 효과적인 감독은 핵심적인 도전 과제가 될 것이다. 이때 중요한 점은, 더 약하지만 신뢰할 수 있는 모니터가 더 강력하고 잠재적으로 신뢰할 수 없는 에이전트를 어떻게 안정적으로 평가할 수 있는가이다. 모델의 내부 추론 과정에 대한 관측 창을 제공하여, CoT 모니터링은 이러한 감독 격차를 줄일 수 있는 특별한 기회를

제공할 수 있다. 실제로 GPT-4o가 특정 보상 해킹 행동에 대해, 더 강력한 모델(OpenAI o3)의 CoT를 효과적으로 모니터링할 수 있음을 발견되었다. 따라서 에이전트와 모니터 간의 연산 자원 및 역량 차이에 따라 모니터링 가능성이 어떻게 확장되는지는 중요한 연구이다.

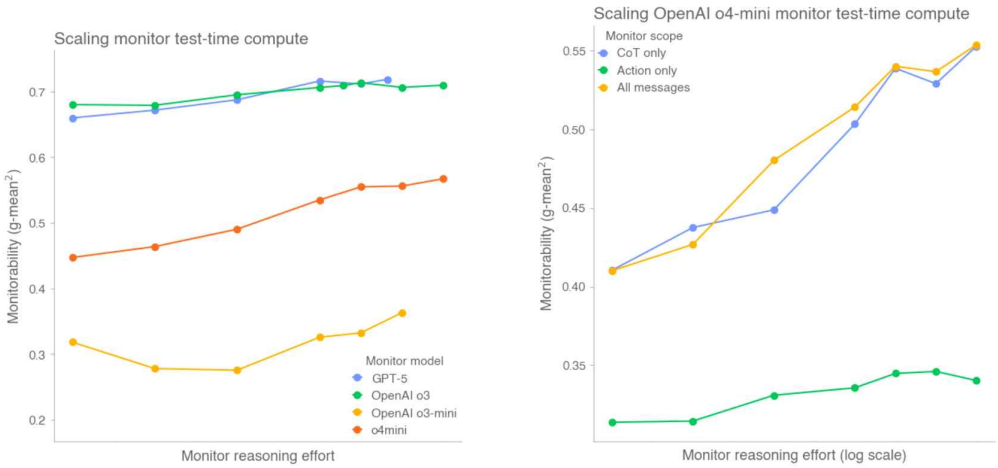


Figure 12: **Effect of Monitor Capacity on Monitorability.** We plot g-mean^2 across our monitorability suite over increasing monitor reasoning efforts. (Left) g-mean^2 with GPT-5 as the agent for multiple monitors over monitor reasoning efforts. All monitors are only given access to the CoT, with GPT-5 and OpenAI o3 being the most capable, followed by OpenAI o4-mini, and OpenAI o3-mini. (Right) g-mean^2 for an o4-mini monitor when given access to different observation scopes, across monitor reasoning efforts.

먼저 GPT-5 Thinking을 에이전트 모델로 설정하고, 서로 다른 모니터 모델들로 모니터링 가능성을 분석하였다(그림 12 참조). 동일한 에이전트를 대상으로 할 때, 더 강력한 모니터 모델일수록 더 높은 모니터링 가능성 점수를 달성함을 확인하였다. 구체적으로, 다양한 추론 노력 설정 전반에서 g-mean² 지표가 o3-mini에서 GPT-5 Thinking로 바뀔 때 약 40% 가까이 증가한다. 또한 추론 모니터의 추론 노력 수준을 확장할수록 모니터링 가능성 역시 증가하는 현상을 관찰했다. 이 효과는 특히 가장 약한 모니터에서 두드러지게 나타난다. 예를 들어 OpenAI o4-mini의 경우, 테스트 시점 연산 자원을 증가시키면 모니터링 가능성이 유의미하게 향상된다. 가장 낮은 추론 노력 설정에서 가장 높은 설정이 될 때, 모니터링 가능성이 약 12% 개선된다. 주목할 점은, OpenAI o4-mini가 에이전트의 행동과 출력만을 제공받는 경우에는, 추론 노력을 증가시켜도 모니터링 가능성이 약 4% 정도만 개선된다는 것이다. 이는 모니터에게 CoT에 대한 접근 권한을 부여할 때, 테스트 시점 연산 자원 증가에 따른 모니터링 가능성 향상 효과가 훨씬 가파르게 나타난다는 것을 의미한다.

7 CoT로부터 더 높은 모니터링 가능성 얻기

일반적으로 CoT 모니터링 가능성은 고정된 값으로 간주된다. 즉, 주어진 모델, 추론 노력 수준, 그리고 특정 프롬프트에 대해 샘플링된 CoT가 있을 때, 해당 샘플의 모니터링 가능성을 개선하는 방법은 모니터 자체를 변경하는 것뿐이다. 그러나 6.3절에서 살펴본 바와 같이, CoT가 길어질수록 모니터링 가능성이 전반적으로 향상되는 경향이 있음을 확인하였다. 이는 모니터에게 제시할 수 있는 토큰을 인위적으로 더 많이 유도할 수 있는 방법이 있는지에 대한 질문으로 이어졌다. 이를 위해, 에이전트가 과제에 대한 초기 응답을 제시한 이후, 문맥 내에서 후속 질문을 추가로 제시하는 실험을 수행한다. 이후, 해당 후속 질문과 그 질문에 대한

모델의 추론에 대응하는 CoT를, 원래의 문제와 CoT와 함께, 모니터에게 제공한다.

실증적으로, 후속 질문은 모니터의 목적에 맞게 구체적이고 맞춤화되어 있을 때 가장 효과적인 것으로 나타났다. 이에 따라, 개입에 대해 직접적으로 질문을 설계하기가 비교적 쉬운 개입 평가에서 g-mean²의 향상이 가장 크게 나타난다. 한편, 모델이 반드시 기만하도록 유인되지 않는 과정 평가의 경우에도, 낮은 추론 노력 수준에서는 g-mean²가 소폭 개선되는 현상이 관찰된다. 이는 후속 질문이 모델에게 자신의 추론을 언어적으로 더 자세히 표현할 수 있는 추가적인 공간을 제공하기 때문일 가능성이 높다(부록 J.2 참조).

이 방법을 실제로 적용하려 할 때, 어떤 표적화된 후속 질문을 던져야 하는지 사전에 알기 어려운 경우가 많다. 물론 각 과제마다 다양한 표적화된 후속 질문들을 사용하는 전략을 고려할 수는 있지만, 이는 전반적으로 적용하기에는 비용이 지나치게 클 가능성이 있다. 이에 대한 자연스러운 대안은, 모든 데이터셋에 공통으로 적용할 수 있는 일반적인 후속 질문을 사용하는 것이다. 일반적인 후속 질문에 대해 단일 응답만 샘플링할 경우, 후속 질문을 전혀 사용하지 않는 기준선 대비 성능 향상은 거의 나타나지 않는다. 반면, 일반적인 후속 질문에 대해 여러 개의 완성(completion)을 샘플링하면, 표적화된 후속 질문을 한 번 샘플링했을 때와 유사한 수준의 효과를 보이며 성능이 크게 향상된다(부록 J.1 참조). 이는 관련된 이상 행동을 사전에 알 수 없는 경우에도, 충분한 샘플링이 주어지면 일반적인 후속 질문이 효과적일 수 있음을 시사한다. 후속 질문에 대한 모델의 최종 답변을 모니터에게 제공하는 경우보다, 그 답변에 이르기까지의 CoT를 제공할 때 모니터링 가능성이 더 높게 나타난다. (4절과 일치).

유망한 접근법이긴 하지만, 모니터링 가능성을 향상시키기 위한 도구로서의 후속 질문에는 한계도 존재한다. 예를 들어 Scruples와 같이, 개입이 명시적인 내용이 아니라 미묘한 프레이밍을 통해 이루어지는 데이터셋의 경우, 의미 있는 맞춤형 후속 질문을 작성하기가 어렵고, 모델은 질문을 받았을 때 자신이 영향을 받았다는 사실을 부인하는 경향을 보인다. 이러한 환경에서는 후속 질문이 오히려 모니터링 가능성을 저하시킬 수 있다. 더 나아가, 후속 질문의 유형이 모델에게 분포 외 입력이 될 경우, 모델이 정보를 제공하기보다는 불규칙하게 행동할 가능성도 있다. (후속 질문 CoT 예시: 부록 J.3)

9 논의 및 결론

CoT의 모니터링 가능성을 추적하고 보존하며 향상시키기 위해서는, 이를 평가할 수 있어야 한다. 본 연구에서는 모니터가 에이전트의 행동과 관련된 특정 속성을 정량적으로 예측하는 양측적(two-sided) 측정지표인 “모니터링 가능성”을 평가하기 위한 대규모 평가 체계를 제안하였다. 기존 연구들 또한 모니터링 가능성을 측정하려는 시도를 해왔지만, 대체로 현실성이 떨어지거나 데이터셋 규모가 제한된 단일한 설정에 국한되어 있었다. 이에 비해 본 연구는 세 가지 서로 다른 평가 원형에 걸쳐 폭넓은 평가 모음을 구축함으로써, 최첨단 추론 모델들이 CoT에서 무엇을 어느 정도까지 언어화하는지를 보다 포괄적으로 측정할 수 있기를 기대한다. 또한, 이미 구축한 평가들을 기반으로 지속적으로 확장해 나가며, 모델의 모니터링 가능성에 대해 더 강력한 보증을 제공하고자 한다.

전반적으로 최첨단 추론 모델들은 비교적 높은 모니터링 가능성을 보였지만, 평가 전반에 걸쳐 성능이 균일하지는 않았다. 특히, 이들 모델은 복잡한 코딩 환경에서의 보상 해킹 의도에 비해, 영합적 행동(sycophantic behavior)을 언어적으로 드러내는 경향이 더 약한 것으로 나

타났다. 본 연구의 평가는 언어화하려는 성향(propensity to verbalize)과 언어화가 필요한 정도(necessity to verbalize)를 구분하여 측정하지 않기 때문에, 영합적 성향이 존재하더라도 그것이 모델 가중치에 내재화되어 있어 굳이 CoT로 표현될 필요가 없을 가능성도 배제할 수 없다. 향후 연구에서는 본 평가 체계를 활용하여, 행동 유형과 환경에 따라 CoT의 필요성이 서로 다른 다양한 상황에서 모니터링 가능성이 어떻게 달라지는지를 보다 잘 이해하고자 한다.

또한, 모니터링 가능성 저하를 유발할 수 있는 몇 가지 잠재적 메커니즘-직접적인 CoT 보상 압력, RL 연산 자원, 그리고 사전학습 연산 자원-을 분석했다. 그러나 조사하지 못한 다른 많은 과정들이 존재하며, 이는 향후 연구 과제로 남겨둔다. 예를 들어, RL 과정에서 CoT에 직접적인 압력을 가해 학습하지는 않지만, CoT 모니터를 사용해 오작동을 탐지하고 취약한 환경을 수정한다. 이때 탐지할 수 있는 환경만을 수정하기 때문에, 장기적으로는 CoT가 점진적으로 난독화되도록 유도하는 느린 진화적 압력이 형성될 수 있다. 또한 숙고적 정렬(deliberative alignment)[19]은 CoT 보상 모델을 상대로 한 단일 단계의 전문가 반복(expert iteration)과 동등하다. 이는 전체 RL 학습 동안 CoT에 직접적인 압력을 가하는 경우에 비해 최적화 압력이 상대적으로 약할 가능성이 높지만, 그럼에도 불구하고 일부 난독화 행동을 유발할 수 있다.

본 연구 결과는 현재의 최첨단 규모에서는 강화학습이 CoT 모니터링 가능성을 유의미하게 저해하지 않는다는 점을 시사하지만, 모델 규모가 더 커질 경우에는 그럴 가능성을 배제할 수 없다. 사전학습 실험에서는, 더 작은 모델에 더 높은 추론 노력(reasoning effort)을 부여함으로써, 더 큰 모델에 낮은 추론 노력을 적용했을 때와 동일한 수준의 역량을 달성하면서도, 모니터링 가능성을 크게 향상시킬 수 있음을 보여주었다. 다만 이러한 향상은 추론 시점 연산 비용을 수반하며, 동일한 추가 연산 자원을 더 큰 모델의 추론 노력을 높이는 데 사용하는 경우와 비교했을 때 역량 측면의 비용(capability cost)도 존재한다. 그럼에도 불구하고, 모델의 역량이 더욱 향상되고 높은 위험도가 수반되는 환경에 배포될수록, 더 높은 모니터링 가능성을 확보하기 위해 이러한 '모니터링 가능성 세금(monitorability tax)'을 지불하는 것이 타당해질 수 있다.

그리고, 후속 질문을 통해 모델로부터 더 많은 사고 토큰(thinking tokens)을 유도하는 매우 기초적인 방법을 실험했다. 이러한 단순한 방법만으로도 모니터링 가능성이 개선될 수 있음을 확인했으며, 향후 연구 공동체가 이를 바탕으로 더 발전된 기법을 제안하길 기대한다.

결론적으로, CoT 모니터링이 확장 가능한 통제(scala-ble control) 체계에서 핵심적인 계층(load-bearing layer)이 될 수 있으며, 점점 더 역량이 높은 모델을 배포하기 위한 안전성 사례(safety cases)에 궁극적으로 이를 통합할 수 있기를 기대한다. 또한 CoT 모니터링을 지속적으로 활용함으로써, 모델이 무엇에 정렬(aligned)되어 있는지, 그리고 모델 자체를 더 깊이 이해할 수 있기를 바란다. 연쇄적 사고 모니터링 가능성을 유지하거나 향상시키기 위해서는 견고하고 폭넓은 평가 세트가 필요하며, 본 연구에서 제안한 평가 체계가 그 방향으로 나아가는 의미 있는 첫걸음이라고 믿는다.

보건 분야 인공지능의 윤리 및 거버넌스: WHO AI 윤리·거버넌스
Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models

https://www.who.int/publications/i/item/9789240084759

취지: 보건 분야에서 LMM(Large Multi-Modal Model) 사용과 관련된 이점과 과제를 파악하고 적절한 개발, 제공 및 사용을 위한 정책과 관행을 수립할 수 있도록 WHO에서 2025년3월 25일에 발간한 가이던스를 살펴본다.

AI는 인간이 매 단계 명시적으로 프로그래밍을 하지 않아도 데이터의 학습에 의해 주어진 과제를 해결한다. 이의 발전으로 생성형 AI가 등장하여 텍스트, 이미지 또는 동영상 같은 새로운 콘텐츠를 생성할 수 있다. 생성형 AI의 한 유형으로 볼 수 있는 LMM(Large Multi-Modal Model)은 하나 이상의 유형으로 데이터를 받아들여 입력된 데이터의 유형에 국한되지 않고 다양한 출력을 생성할 수 있기에 보건의료, 과학 연구, 공중보건 및 신약 개발에 널리 사용될 것으로 여겨진다.

2021년 WHO는 “보건 분야에서 AI의 윤리와 거버넌스에 관한 포괄적인 가이던스”를 발표하여, 정부, 공공 부문 기관, 연구자, 기업, 실행자 등을 포함한 폭넓은 이해관계자들이 보건의료 분야에서 AI를 개발하고 배포하는 데 있어 지켜야 할 6가지 원칙을 다음과 같이 도출하였다: (1) 자율성 보호, (2) 인간의 복지, 안전 및 공익 증진, (3) 투명성, 설명 가능성 및 이해 가능성 보장, (4) 책임감과 책무성 강화, (5) 포용성과 형평성 보장, (6) 반응성이 높고 지속 가능한 AI 촉진.

그 후속으로 WHO 회원국들이 보건 분야에서 LMM 사용과 관련된 이점과 과제를 파악하고 적절한 개발, 제공 및 사용을 위한 정책과 관행을 수립할 수 있도록 가이던스를 2025.3.25.에 발행하였다. 이 가이던스에는 원칙에 부합하는 기업 내 거버넌스, 정부의 역할, 국제 협력을 통한 거버넌스 권고사항이 포함되어 있으며, 주요 내용은 다음과 같다.

• LMM의 응용, 과제 및 위험:

- 보건의료 분야에서 LLM의 다양한 이점 및 위험성
- LMM 사용과 관련된 보건 시스템의 위험성
- LLM 사용 증가와 관련된 더 광범위한 사회적 위험과 탄소 배출 및 물 부족

• 보건의료 및 의학 분야에서 LMM의 윤리 및 거버넌스:

- 범용 기반 모델(LMM) 설계 및 개발
- 범용 기반 모델(LMM) 서비스 제공
- 범용 기반 모델(LMM) 배포
- LMM에 대한 법적 책임
- LMM의 국제 거버넌스

1 서론

본 가이던스는 보건 관련 응용 분야에서 LMM의 사용에 대해 설명한다. 여기에는 보건의료 및 의학 분야에서 LMM 사용의 잠재적 이점과 위험, 그리고 윤리, 인권 및 안전에 관한 가이던스와 의무를 준수하도록 보장하기 위한 LMM 거버넌스 접근방식이 포함된다. 본 가이던스는 WHO가 2021년 6월에 다음 내용으로 발행한 “보건 분야 인공지능의 윤리 및 거버넌스”에 기반하고 있다: “보건 분야 AI 사용의 윤리적 도전과 위험”, “AI가 모든 국가의 공익을 위해 사용되도록 보장하는 여섯 가지 원칙”, “기술의 잠재력을 극대화하기 위해 보건 분야 AI 거버넌스를 강화하기 위한 권고사항”

LMM은 보건 및 의학 분야에서 사용하기 위해 훈련되며, 학습 데이터에는 텍스트를 넘어 바이오센서, 게놈, 에피게놈, 프로테오믹스, 영상, 임상, 사회적, 환경적 데이터 같은 매우 다양한 데이터가 포함된다. LMM은 입력된 데이터 유형에 국한되지 않는 출력을 생성할 수 있기에 보건 분야와 신약 개발에서 다양한 응용을 위한 모델로 기대되고 있다.

AI 가치사슬은 주로 대규모 기술 기업에서 시작된다. 개발자는 LMM을 개발하는 데 필요한 “AI 인프라”를 구성한다. 일부 LMM은 보건 및 의학 분야에서 사용하기 위해 학습되었다. LMM은 제 3자(“제공자”)가 활성화된 프로그래밍 인터페이스를 통해 특정 목적 또는 용도로 사용할 수 있다. 그 이후, 제공자는 LMM을 기반으로 한 제품이나 서비스를 보건부, 보건의료 시스템, 병원, 제약회사, 또는 의료 제공자와 같은 고객에게 마케팅하며, 제품이나 응용 프로그램을 구매하거나 라이선스를 취득한 고객은 이를 환자, 의료 제공자, 보건의료 시스템 내의 다른 단체, 일반 대중, 또는 자신의 사업에서 직접 사용할 수 있다.

WHO는 AI가 공중보건 개선과 보편적 건강 보장 달성을 포함하여 보건 시스템에 제공할 수 있는 막대한 이점을 인식하고 있지만, “보건 분야 AI의 윤리 및 거버넌스에 대한 가이던스”에서 설명된 바와 같이, AI는 공중보건을 약화 시키고 개인의 존엄성, 개인정보, 인권을 위태롭게 할 수 있는 중대한 위험을 수반하고 있다. 따라서, WHO는 급속하게 사용이 확산되고 있는 LMM이 전 세계적으로 성공적이고 지속적으로 사용될 수 있도록 가이던스를 제공한다.

1.1 LMM의 중요성

LMM은 비교적 새로우며, 아직 검증되지 않았음에도 불구하고 보건의료 및 의학 분야를 포함한 다양한 영역에서 사회에 엄청난 영향을 미치고 있다. 현재 많은 기업이 LMM을 개발하거나 인터넷 검색 엔진과 같은 소비자 응용 프로그램에 통합하고 있다. LMM의 등장이 기술 산업에 새로운 투자를 촉진하고 신제품들이 출시되는 가운데, 일부 기업은 LMM이 특정 응답을 생성하는 이유를 완전히 이해하지 못한다고 인정하고 있다. 인간 피드백을 활용한 강화 학습에도 불구하고, LMM은 항상 예측 가능하거나 통제 가능한 결과를 생성하지는 않는다.

LMM은 빠르게 도입되었지만, 이를 학습시키는 데 사용된 데이터는 공개되지 않아서 데이터 편향, 합법적 수집, 데이터 보호 규정 및 원칙 준수, 그리고 특정 작업이나 질문에 대한 성과가 동일하거나 유사한 문제로 학습된 결과인지 아니면 문제를 해결하는 능력을 습득한 결과인지를 알기 어렵다. 더군다나 LMM의 활용이 개인과 정부의 준비가 충분하지 않은 상태에서 급속도로 번지고 있다. 개인은 LMM을 효과적으로 사용하는 방법에 대해 교육받지 못했으며, LMM이 생성하는 응답이 항상 신뢰할 수 있는 것이 아니라는 사실을 이해하지 못할 수 있다. 정부 역시 대부분 준비가 부족했다. AI 사용을 규제하기 위한 법률과 규정은 LMM과 관련된

과제나 기회를 다룰 준비가 되어 있지 않다. 앞으로 더 강력한 LMM이 계속 출시될 것으로 예상되며, 이는 새로운 혜택뿐만 아니라 새로운 규제 과제도 가져올 것이다. 이러한 역동적인 환경에서, 이전의 윤리적 가이던스를 포함한 가이던스를 바탕으로 보건 및 의료 분야에서 LMM을 사용하는 데 대한 제안과 권장 사항이 필요하다.

1.2 보건분야를 위한 AI 윤리 및 거버넌스에 대한 WHO 가이던스

WHO의 "보건분야 AI 윤리 및 거버넌스 가이던스"(2021)는 다양한 머신러닝 접근방식과 보건 의료에서의 AI의 다양한 응용을 검토했지만, 생성형 AI나 LMM에 대해 구체적으로 다루지는 않았다. 2021년에 가이던스를 발표할 당시에도, 생성형 AI와 LMM이 이렇게 빠르게 보급되고 임상 진료, 보건 연구, 공중보건에 적용될 것이라고 예측하지 못했다.

그럼에도 불구하고, 2021년의 가이던스에서 파악된 근본적인 윤리적 도전과제와 핵심 윤리 원칙 및 권고사항(Box 1)은 여전히 유효하며, LMM을 평가하고 효과적이고 안전하게 사용하는 데 적용될 수 있으므로, 이 가이던스에 제시된 LMM에 대한 접근의 기초로 활용되었다.

Box 1. WHO 합의에 기반한 보건분야 AI 사용을 위한 윤리 원칙 개요

- **자율성 보호:** 인간은 보건의료 시스템과 의료 결정에 대한 통제권을 유지해야 한다.
- **인간의 복지, 안전 및 공익 증진:** AI 설계자는 명확하게 정의된 용도 또는 적응증에 대해 안전성, 정확성 및 유효성에 대한 규제 요구사항을 충족해야 한다.
- **투명성, 설명 가능성 및 이해 가능성 보장:** AI 기술은 개발자, 의료 전문가, 환자, 사용자 및 규제기관이 이해할 수 있도록 설계되어야 한다.
- **책임감과 책무성 강화:** AI는 적절한 조건에서 적절히 훈련된 사람들에 의해 사용되도록 책임성을 강화해야 한다.
- **포용성과 형평성 보장:** AI는 연령/성별/성정체성/소득/인종/민족/성적 지향/역량 등에 상관없이 널리, 적절히, 공평하게 사용되고 접근할 수 있도록 설계되고 공유되어야 한다.
- **반응성이 높고 지속 가능한 AI 촉진:** AI 기술은 보건시스템, 환경, 직장의 지속 가능성을 촉진하는 방향과 일치해야 한다.

I LMM의 응용, 과제 및 위험

2 보건 분야에서 LMM 사용의 응용 및 과제

AI의 보건 분야 응용은 진단, 임상 진료, 연구, 약물 개발, 보건의료 행정, 공중보건 및 감시 등이 있다. LMM 응용은 임상의, 환자, 일반인, 보건의료 전문가 및 종사자에 따라 다르게 사용된다. 여기서는 보건의료 분야에서 LMM의 잠재적 응용과 과제 및 위험에 대해 다루어본다.

2.1 진단 및 임상진료

AI는 방사선/의료영상, 결핵/종양학 분야에서 진단과 임상 진료를 지원하고 있다. 임상의는 상담 중 환자기록 통합, 위험 환자 식별, 어려운 치료 결정, 임상적 오류 포착에 AI의 도움을 받고 있다. LMM은 진단 및 임상 진료 전반으로 AI 시스템 사용을 확장시키며, "과거의 청진기보다 의사들에게 더 중요한 역할을 할 것"으로 기대된다. LMM은 미국 의학 면허 시험을 통과했지만, 의료지식의 단순 반복으로 시험을 통과하는 것은 안전하고 효과적인 임상 서비스와 다르다. "LMM이 의료 질문 답변에서 보건의료 제공자/관리자/소비자가 사용할 도구로 전환되려면, 기술의 안전성/신뢰성/유효성 및 개인정보 보장에 대한 추가 연구가 필요하다."

진단은 특히 유망한 분야로, LMM은 복잡한 사례에서 희귀 진단이나 비정상적 증상을 식별하며, 진단을 위한 추가 도구가 될 것이다. 또한, 명백한 진단이 간과되지 않도록 의사에게 추가적인 의견을 제공한다. 이는 LMM이 환자 의료기록을 의사보다 훨씬 빨리 스캔하기 때문이다. 현재 보건 데이터로 특정하여 훈련되지 않은 LMM이 임상의를 지원하기 위한 파일럿 프로그램에서 사용되고 있다. 이는 의료진이 환자 문의에 답변하거나 수천 건의 메시지를 처리하는 데 따른 번아웃을 줄여 임상업무에 집중할 수 있게 한다. 한 연구에서 ChatGPT 기반 챗봇이 온라인 포럼 질문에서 의사보다 더 나은 성과를 보였다. 챗봇은 표준화된 "비공식적 상담" 질문답변과 초기 진단 단계에서 정보와 응답을 제공하고 검사 결과를 요약하는 데도 유용하다.

기업과 대학들은 의료/보건 데이터 또는 전자 건강 기록을 활용해 훈련된 LMM을 개발하고 있다. 몇몇 LMM은 수백만 개의 전자 건강 기록과 기타 전문 및 일반 의학 지식을 포함한 소스 데이터를 기반으로 알고리즘에 의해 훈련되어 "의료 질의 응답" 능력을 향상시켰다. 몇몇 대규모 기업들은 일반 용도의 LMM을 임상진단 및 치료를 도와주는 모델로 변환시켜 의료 종사자와 컴퓨터 간의 "의견 불일치"를 완화하도록 하고 있다.

장기적인 비전은 "범용 의료 인공지능"을 개발로 보건의료 종사자들이 LMM과 유연하게 대화하여 맞춤형 임상의 주도 질의에 따라 응답을 생성하게 하는 것이다. 이를 통해 사용자는 일반적인 언어로 요구사항을 설명하여 범용 의료 AI 모델을 새로운 작업에 적용할 수 있다.

진단 및 임상진료에서 LMM 사용의 다섯 가지 주요 위험

• **부정확하거나 불완전하고 편향적이거나 잘못된 응답:** 챗봇이 데이터를 기반으로 잘못되거나 완전 허위 응답을 생성하고, 학습 데이터의 결함으로 편향된 응답을 생성할 수 있다. LMM은 또한 사용 환경에 대한 가정을 기반으로 다른 환경에 적합한 권장 사항을 제공하는 맥락적 편향을 초래한다. 예) 저/중소득 국가 데이터가 없는 경우, 저소득 국가의 질병 치료 방안 요청에 고소득 국가의 방식을 제공할 수 있다. 또한, 사용 환경의 변화된 상황을 반영하지 못한 응답을 제공할 수 있다. 소위 "환각(hallucination)"으로 알려진 응답은 정확한 응답과 구별할 수 없다. 이는 LMM이 인간 피드백을 통한 강화 학습을 거쳤더라도, 사실처럼 보이는 정보를 생성하도록 훈련되기 때문이다. LMM은 또한 "프롬프트 엔지니어링"이라는 인간의 입력 최적화에 의존한다. 따라서 의료 데이터와 보건 정보를 기반으로 훈련된 LMM이 반드시 올바른 응답을 생성한다고 보장할 수 없다. 보건의료/공중보건 의사결정 영역에서, LMM을 안전하고 효과적으로 보건의료 시스템에 구현/개발하는 비용을 정당화할 만큼 정확하지 않을 수 있다.

• **데이터 품질 및 데이터 편향:** LMM이 편향되거나 부정확한 응답을 생성하는 주요 원인 중 하나는 데이터 품질 때문이다. 인터넷과 같은 대규모 데이터 세트를 기반으로 훈련되었는데, 이 데이터 세트는 잘못된 정보와 편향으로 가득하다. 대부분의 의료 및 보건 데이터 역시 인종, 민족, 조상, 성별, 성 정체성 또는 연령에 관계없이 편향되어 있다. 특히 대부분의 데이터가 고소득 국가에서 수집된다. LMM은 또한 오류와 부정확한 정보로 가득 찬 전자 건강 기록으로 훈련되거나, 신체 검사에서 얻은 부정확한 정보에 의존할 수 있다. 데이터 품질 및 편향 문제는 LMM을 포함한 모든 AI 모델에 영향을 미친다. LMM은 알고리즘이 학습된 데이터의 종료 시점에 의해 제한될 수 있지만, 일부 LMM은 이제 인터넷에서 최신 정보를 검색할 수 있다. 이는 더 많은 잘못된 정보나 부정확한 정보를 생성할 가능성을 증가시킬 수 있다. 의학에서 최신 정보와 높은 정확성이 표준 진료를 충족시키고 특정 질병을 이해하는 데 중요하다.

• **자동화 편향:** LMM이 보건의료 전문가에게 자동화 편향을 조장할 가능성 때문에 우려가 더

욱 커진다. 자동화 편향이란, 임상의가 사람이 발견하였어야 할 오류를 간과하는 것을 말한다. 또한 의사/보건의료 종사자가 윤리적 또는 도덕적 사항이 상충되는 결정을 내릴 때 LMM을 사용할 수 있다는 우려도 있다. LMM을 도덕적 판단에 사용하면 의사들이 어려운 판단이나 결정을 내리지 못하게 되는 "도덕적 탈숙련화, 즉 도덕적 역량 상실"로 이어질 수 있다.

• **기술 저하:** 의료행위에서 AI 사용이 증가함에 따라 의사들이 점점 더 일상적인 책임과 의무를 컴퓨터에 의존하게 되면서, 장기적으로 의료 전문가로서의 역량 저하/약화 위험이 있다. 기술의 상실은 의사가 AI 알고리즘의 결정을 자신 있게 무효화 하거나 도전하지 못하게 하고, 네트워크 장애나 보안 침해 시 의사가 특정 의료 작업과 절차를 완료하지 못할 수 있다.

• **사전동의:** LMM의 사용 증가로, 환자에게 AI 기술이 응답을 지원하거나 임상의의 피드백을 모방한 응답을 생성할 수 있다는 사실을 알려야 한다. 그러나 LMM/AI가 일반적 의료행위에 통합되면, 환자/보호자는 AI 기술에 전적/부분적으로 의존하는 것이 불편하거나 원하지 않더라도, 사용에 대한 동의를 거부할 수 없게 된다. 이는 AI에 기반하지 않는 다른 옵션을 쉽게 이용할 수 없거나, 임상의가 AI를 사용하지 않고 의료 서비스를 제공할 수 없는 것이다.

2.2 환자 중심의 응용

환자들은 약물 복용/영양과 식단 개선/신체 활동 참여/상처 관리/주사 투여 등의 건강 관리 활동을 한다. AI는 이러한 자가 관리의 범위를 확대할 것이다.

LMM은 환자와 일반 대중이 의료 목적으로 AI를 사용하는 추세를 가속 시킨다. LMM은 인터넷 검색에 통합되어 환자와 대중에게 정보를 제공할 것이다. LLM 기반 챗봇은 자가 진단 및 병원 방문 전 정보를 찾는 데 있어 검색 엔진을 대체할 것이며, 다양한 형태의 데이터를 활용하여 매우 개인화되고 폭넓은 초점의 가상 건강 비서로 활용될 수 있다. "가상 건강 조언자는 개인 프로필을 활용하여 행동 변화를 촉진하고, 건강 관련 질문에 답하며, 증상을 분류하거나 적절한 경우 의료 제공자와 소통할 수 있다"고 한다. 환자 중심 LMM의 세 번째 응용 분야는 임상 시험을 식별하고 시험에 참여하는 것이다. AI 기반 프로그램이 환자와 임상 시험 연구자가 적합한 시험을 찾는 것을 돕고 있는데, LMM은 환자의 의료 데이터를 활용하여 동일한 방식으로 사용될 수 있다. 이는 모집 비용을 줄이고 속도와 효율성을 높이며, 다른 채널을 통해 찾기 어렵고 접근하기 힘든 적절한 시험과 치료 기회를 개인에게 더 많이 제공할 수 있다.

위험과 도전 과제: 개인이 LMM을 쉽게 사용할 수 있다는 점이 지닌 중대한 위험

• **부정확/불완전하거나 잘못된 정보:** 환자 및 일반인도 LMM 혹은 의료 정보 제공 AI 프로그램 사용 시 부정확/불완전하거나 편향되고 잘못된 진술의 위험이 있다. 의료 전문 지식이 없는 사용자가 LMM을 사용할 경우, 응답을 검토하거나 반박할 근거가 없거나, 다른 정보 출처에 접근할 수 없는 상황, 또는 어린이가 이를 사용하는 경우 이러한 위험이 더욱 커진다.

• **조작:** 많은 LMM 기반 챗봇 프로그램은 대화 방식에서 서로 다른 접근을 취하며, 점점 더 설득력 있고 중독성이 강해질 것이다. 챗봇은 사용자에 맞춘 대화 패턴을 채택하며, 질문에 답변하거나 대화에 참여하여 개인이 자신에게 해로운 행동(자살)을 하도록 유도할 수 있다. 전문가들은 챗봇이 "감정적으로 조작적"이 될 가능성이 있다며 긴급 조치를 촉구하고 있다.

• **개인정보:** LMM 사용 시 식별 가능한 의료 정보가 LMM에 입력 후 제 3자에게 공개될 가능성이 있다. LMM에 공유된 데이터가 반드시 삭제되는 것은 아니며, AI 모델 개선을 위해 사용될 수 있다. 다른 사용자가 LMM을 통해 특정 정보를 요청하거나, 한 LMM을 여러 사람과 공유할 경우 다른 사람들의 기록이 잘못 공개되는 경우가 발생할 수 있다.

• **임상의와 환자/일반인의 상호작용 약화:** 환자/보호자의 LMM 사용은 의사-환자 관계를 근본적으로 변화시킨다. 환자들이 인터넷 검색으로 얻은 의료 정보를 바탕으로 의료 제공자에게 도전하거나 추가 정보를 요구할 수 있다. LMM은 이러한 대화를 개선할 수 있지만, 환자/보호자가 LMM에 의존하여 결정함으로써 의학적 판단/지원에 의존하지 않게 된다. AI 기술이 임상의와의 접촉을 줄이면, 건강증진 도모 기회가 줄어들고 환자와의 상호작용이 약화되어 지지의료가 약화될 수 있다. 일반적으로, AI로 인해 임상진료가 "비인간화"될 수 있다.
• **인식적 부정의:** 보건의료 제공자의 판단을 LMM의 판단으로 대체할 경우 환자에게 "인식적 부정의"를 초래할 수 있다. "인식적 부정의"란 "지식의 주체로서의 특정 사람에게 가해지는 부당함"을 의미하며, 이는 의료 시스템 내에서 환자에게 발생할 수 있다. 인식적 부정의의 한 형태인 해석적 부정의는 공유된 이해와 지식(소위 "집단적 해석 자원")의 격차로 인해 일부 사람들이 생활 경험, 사회 경험, 자신의 신체적/정신적 상태를 이해하는 데 불이익을 받게 되는 상황을 말한다. LMM은 대규모 데이터에 기반해 학습되었더라도 인식하고 반응할 수 있는 범위, 즉 자신이 가진 어휘를 초과하는 개념과 관념에 한계가 있다. 만약 환자의 경험이 임상 환경에서 LMM에 의해 인정되지 않는다면, 이는 의료 제공자의 적절한 치료를 방해할 수 있다. 이러한 상황은 데이터에서 소외되고 대표성이 부족한 취약 계층에서 자주 발생할 것이다.
• **의료 시스템 외부에서의 보건의료 제공:** 건강 관련 AI 응용 프로그램은 더 이상 의료 시스템 내에서의 독점적 사용을 넘어서, 비의료 시스템에서도 쉽게 획득하여 사용되거나 도입되고 있다. 이제 이러한 기술이 더 높은 규제 감독이 요구되는 임상 응용 프로그램으로 규정해야 할지, 아니면 "웰니스"와 같은 비교적 낮은 규제와 검토가 요구되는 응용 프로그램으로 규정해야 할지 의문이다. 현재 이 기술은 두 범주 사이의 회색 지대에 속한다고 여겨진다.

환자가 규제 안전장치 없이 LMM을 의학적 조언이나 자가 진단에 사용하면 위험을 초래한다. 환자가 잘못되거나 오해를 불러일으킬 수 있는 조언을 받을 수 있으며, 특히 자살 충동을 가진 개인이 AI 챗봇을 사용할 때 환자 안전이 위협받을 수 있다. 설령 정보가 정확하더라도, 의료 교육을 받지 않은 개인이 해당 정보를 잘못 해석하거나 오용할 수 있다. 이러한 응용 프로그램이 계속 확산되지만 의료 응용 프로그램으로만 분류되는 것은 아니기 때문에 전반적인 의료 서비스의 질이 저하될 가능성이 있다. 특히 다른 대안이 부족한 사람들이 이러한 응용 프로그램에 의존하게 되면 양질의 의료 서비스에 대한 접근성이 더욱 불평등해질 수 있다.

2.3 사무 기능 및 행정업무

의학 분야에서 환자 정보/데이터의 전자 건강 기록, 민간/보험/공공 의료시스템의 청구서 처리, 기타 행정 업무 수행으로 의사와 의료 종사자들의 번아웃 되고 있다. LMM은 보건의료 전문가들에게 가장 귀중한 자원인 시간을 되돌려주는 수단으로 주목받고 있으며, 이는 번아웃을 줄이고 환자 치료에 시간을 더 할당하고 더 많은 환자를 볼 수 있도록 한다.

LMM의 현재 및 추후 사용이 예상되는 사례는 다음과 같다: ① 의료 용어를 단순화하고 의사소통을 더 "환자 친화적"으로 만들어 번역 또는 임상의-환자간의 의사소통 개선에 도움을 줌; ② 전자 건강 기록에서 누락된 정보를 채우는 데 활용; ③ 기타 AI 형태와 함께 환자 방문 후(가상 또는 대면) 임상 기록 초안을 작성; ④ 자동 처방전 작성, 예약 관리, 검사 일정 조율, 청구 코드, 보험사의 사전 승인 처리, 시술 노트 및 퇴원 요약서 작성 등의 작업을 사전에 수행; ⑤ 더 정교한 LMM이 개발됨에 따라, 방사선 전문의와 같은 복잡한 문서 작업에도 활용.

위험과 도전 과제: LMM의 부정확성, 실수(예: 필사, 번역 또는 단순화) 또는 "환각"으로 인해 심각한 오류가 발생할 수 있기에 대부분의 사무 및 행정 기능을 완전히 자동화하지 않는 것이 중요하다. 또한 LMM이 일관되지 않을 수 있어 프롬프트나 질문을 약간 변경하면 완전히 다른 전자 건강 기록이 생성될 수 있지만 이러한 불일치는 시간이 지남에 따라 감소할 것이다.

2.4 의학 및 간호교육

LMM은 학생의 특정 요구와 질문에 맞춰 맞춤형으로 "동적 텍스트"를 생성하는 데 사용될 수 있다. 챗봇에 통합된 LMM은 임상의-환자 간의 의사소통 및 문제 해결 능력을 향상시키기 위해 모의 대화를 제공할 수 있으며, 이는 의료 면담, 진단적 추론 및 치료 옵션 설명 연습에 활용될 수 있다. 또한 챗봇은 장애가 있는 환자나 희귀 질환을 가진 환자를 포함한 다양한 가상의 환자를 학생들에게 제공할 수 있다. LMM은 또한 "연쇄적 사고"를 통해 생리학적 및 생물학적 과정을 설명하며 의대생의 질문에 답변하는 형태로 교육을 제공할 수 있다.

위험과 도전 과제: AI 사용은 보건의료 전문가의 훈련과 기술을 향상시킬 수 있지만, 동시에 전문가가 자신의 판단(또는 동료의 판단)을 컴퓨터에 의존하는 위험을 초래할 수 있다. LMM이 부정확한 정보나 응답을 제공 혹은 조작할 경우, 의료 교육의 질에 영향을 미칠 수 있다. 또 다른 우려는 교육에서 LMM의 사용 또는 행정 및 사무 기능을 단순화하는 과정이 디지털 문해력이 부족한 의료 종사자들에게 추가적인 부담을 줄 수 있다는 점이다. 이들은 일상 업무에서 AI 지원 기술 사용 역량을 개발해야 할 필요가 있다. LMM의 새로운 기능은 보건의료 전문가가 지속적으로 재교육받아야 할 것이다. 개발자들은 궁극적으로 일반인도 쉽게 사용할 수 있는 자연어/시각 커뮤니케이션 인터페이스를 갖춘 AI 지원 기술을 도입해야 한다.

2.5 과학/의학 연구와 약물 개발

AI는 과학 및 임상 연구와 신약 개발에 사용되고 있다. AI를 통해 전자 건강 기록을 분석하여 임상 실무 패턴을 식별하고 새로운 임상 실무 모델을 개발할 수 있다. 또한 질병에 대한 이해를 향상시키고 새로운 바이오마커를 식별하는 등 유전체학 분야에서 사용된다. AI는 신약 개발의 거의 모든 단계에서 활용되며, 화합물 선별 과정을 간소화하고, 단백질의 3차원 구조를 예측하며, 전임상 개발 단계에서 화합물의 독성과 효과를 예측하고, 임상 시험 중 참가자 모집, 등록 및 모니터링을 개선하는 데 사용된다.

LMM은 과학/의학 연구방법에도 변화를 주고 있다. 과학 논문 작성, 논문 제출 또는 동료 검토 작성에 사용할 문장을 생성한다. 학술논문요약/초록생성을 하고, 데이터 분석/요약하여 임상/과학 연구에서 새로운 통찰을 얻도록 활용된다. 문장을 편집하여 논문/연구계획서 등 문서의 문법/가독성/간결성을 개선한다. LMM은 신약 발견과 새로운 화합물을 설계하는 de novo 신약 설계에도 사용된다. 여기에는 특정 속성을 가진 화합물을 개발하는 작업이 포함된다.

위험과 도전 과제: 한 학술 출판사의 두 가지 규칙-(1) 연구 논문의 저자로 LMM을 인정하지 않으며, (2) LMM을 사용하는 연구자는 방법론 섹션과 감사의 글에 그 사용을 명시해야 한다. 과학 연구에서 LMM의 사용과 관련된 일반적인 우려는 다음과 같다:

- **책임 부족:** 과학 또는 의학 연구 논문의 저자는 책임성을 가져야 하지만, AI 도구는 이를 가질 수 없다. 이 책임감 부족은 주요 학술 출판사와 세계 의학편집자협회가 LMM을 저자로 인정하지 않기로 한 결정한 근거이다.
- **고소득 국가 편향:** LMM을 학습시키는 데 사용되는 과학 및 의학 연구 대부분이 고소득 국

가에서 수행되기에 LMM의 쿼리 결과는 고소득 국가의 관점을 편향적으로 반영할 가능성이 높다. 이는 저/중소득 국가에서 나온 연구를 무시하거나 인용하지 않는 경향이 강하다.
• **환각 또는 잘못된 정보:** 존재하지 않는 정보의 인용이나 요약으로 "환각"을 일으킬 수 있다.
• **신뢰 저하:** 동료 검토(review) 생성 활동에 LMM을 사용하면 신뢰를 약화시킬 수 있다.
• **LMM 및 LMM이 생성한 지식에 대한 접근성:** LMM의 유료화는 디지털 지식 격차를 악화시키며, 재정적 지원이 부족한 과학자들에게 부정적인 영향을 미칠 수 있다.

AI(LMM 포함)는 신약 개발에 혜택을 제공할 수 있지만, 이 분야에서 AI 사용에 대한 우려도 존재하며, 이는 후속 WHO 출판물에서 검토될 예정이다.

3 보건 시스템 및 사회에 대한 위험과 LMM 사용에 대한 윤리적 문제

의료 분야에서 LMM 및 AI 기술의 사용과 관련된 위험은 (1)국가의 보건 시스템에 영향을 미치는 위험, (2)규제 및 거버넌스와 관련된 위험, (3)국제적인 사회적 우려의 범위로 나뉜다.

3.1 보건 시스템에 대한 영향

보건 시스템은 서비스 제공, 의료인력, 보건 정보 시스템, 필수 의약품 접근성, 재정, 리더십 및 거버넌스의 여섯 가지 구성요소를 기반으로 한다. LMM은 이러한 구성요소에 직간접적으로 영향을 미칠 수 있으며, 이와 관련된 위험은 다음과 같다.
LMM 이점의 과대평가 및 위험의 간과: AI의 가능성을 과장/과대평가하여, 안전성/유효성의 철저한 검증이 되지 않은 제품/서비스의 채택될 수 있다. 이는 "기술적 해결주의" 때문이며, AI/LMM과 같은 기술이 실제 유용하고 안전하며 효과적이라는 것이 입증되기 전에 사회적/구조적/경제적/제도적 장벽을 제거하는 만능 해결책으로 간주 된다. LMM은 정책 입안자/제공자/환자가 그 이점을 과대평가하고 LMM이 초래할 수 있는 문제/과제를 간과하게 만들 수 있다. 정책 입안자는 LMM이 개발되고 사용되기 전까지 이의 사용범위를 정하는데 필요한 증거를 확보하는 것이 어렵다. LMM의 사용이 이미 사용 중인 AI 기술이나 치료적/공중 보건적인 이점이 입증된 비 AI 솔루션 보다 우선 시 되면 안 된다. 특히 LMM에 대한 비균형적인 의료 정책과 잘못된 투자는 주의를 분산시키고 검증된 분야의 자원을 빼앗아 갈 수 있다.
접근성과 경제성: LMM의 공평한 접근을 저해하는 첫째 요인은 디지털 격차이다. 이는 AI 사용에 영향을 미치는 다른 불평등을 초래하며, AI 자체가 이러한 불평등을 강화하고 악화시킨다. 그 다음 요인은 LMM이 사용료/구독료를 지불해야 한다는 점이다. 이는 저/중소득 국가와 자원이 부족한 개인/의료 시스템/지방 정부에서도 특정 LMM을 이용하기 어렵게 한다. 한편, 경제적으로 어려운 사람들은 "가성비 좋은 해결책"으로 LMM에 의존하게 되어 "진짜" 보건의료 전문가에게 접근하는 것은 부유한 사람들만의 특권이 될 가능성이 있다. 세 번째 요인은 대부분의 LMM이 현재 영어로 운영되기에, 다른 언어로 입력을 받고 출력할 수는 있지만 잘못된 정보나 허위 정보를 생성할 가능성이 높아진다.
시스템 전반의 편향: AI 학습 데이터 세트는 여성/소수인종/노인/농촌지역사회/취약계층을 배제하는 등 편향되는 경우가 많다. 일반적으로 AI는 많은 데이터 위주로 학습하기 때문에, 소수를 불리하게 한다. 이러한 편향은 의료 시스템 전반에 걸쳐 차별을 초래하며, 사람들의 보건 서비스 및 고품질 치료에 대한 접근에 영향을 미친다. 동시에, LMM은 편향과 고정관념에 "반발"하는 데이터를 포함할 가능성도 있다. 연구자들은 모델이 고정관념에 의존하지 않도록 프롬프트를 설정했을 때 알고리즘의 응답이 긍정적으로 변하는 것을 보였다.
노동과 고용에 미치는 영향: OECD는 AI기반 자동화로 고도 숙련 직업이 갑작스러운 위험에

처할 수 있다고 하였지만, 의료는 정부의 핵심 기능이기에 의료가 기술로 대체되지 않을 것이고, 많은 국가들은 보건의료 종사자 부족 문제를 겪고 있다. 따라서 안전성/유효성이 입증된 LMM은 필요 의료 인력과 실제 의료 인력 간의 격차를 좁혀줄 것이다. 또 다른 우려는 LMM의 도입이 현재 및 미래의 보건의료 전문가 수에 미칠 영향이다. 컨설팅 회사 액센츄어는 LMM이 작업 시간의 40%에 영향을 미칠 수 있다고 추정하며, “인간의 창의성/생산성에 미칠 긍정적 영향이 막대할 것”이라고 평가했다. 그러나, LMM의 도입은 많은 의료 종사자에게 도전 과제를 초래하며, 이들은 LMM 교육과 적응이 필요하다. 세 번째 우려는 LMM 학습 데이터를 검토하고 학대/폭력/정신적고통 내용의 콘텐츠를 제거하는 책임을 맡은 사람들에게 가해지는 정신적/심리적 부담이다. 이들은 종종 저/중소득 국가에 기반을 두고 있으며, 저임금을 받고 일하면서 상담/의료 서비스에 접근하지 못한 채 심리적 고통을 겪을 가능성이 있다.
부적합한 LMM에 대한 보건 시스템의 의존: LMM은 의료인 부족 문제를 완화하고 보건 시스템의 범위를 확장할 수 있게 해주지만, 보건 시스템이 LMM에 과도하게 의존할 수도 있다. 의료/공중보건용 LMM이 유지되지 않거나, 고소득 국가위주로 설계/업데이트될 경우, LMM에 의존하던 보건 시스템은 LMM 없이 의료 서비스를 제공해야 할 수도 있다. 의료 전문가들이 “기술숙련 저하”를 겪고 특정 책임을 AI에 의존하거나, 환자들이 AI 사용을 기대하는 경우, LMM에 대한 과도한 의존이 의료 시스템에 대한 신뢰를 악화시킬 수 있다. 또 한, 환자의 개인정보/기밀성 보장이 안되면, 개인정보 침해 위험 없이 접근할 수 있다는 확신이 저하된다.
사이버 보안 위험: 보건의료 시스템이 악의적인 공격과 해킹의 표적이 되며, 학습데이터가 조작되어 성능/권장사항 등이 변경될 수 있고, 데이터가 랜섬웨어로 “탈취”될 수도 있다. 특히 민감한 데이터가 무단 공개/사용으로부터 보호되지 않는 LMM에 입력되면 위험하다. LMM 자체도 제 3자가 LMM에 데이터를 입력하여 개발자가 의도하지 않은 방식으로 작동하게 만들 수 있다. 이러한 Prompt injection은 보안 연구자들에 의해 LMM의 과제를 설명하는 데 사용되고 있지만, 악의적인 행위자가 데이터를 훔치거나 사용자를 속이는 데 악용할 수 있다.

3.2 규제 및 법적 요구사항 준수

AI 사용 규제를 위한 기존 데이터보호법 및 국제인권의무는 LMM의 개발/제공/배포에 적용될 수 있다. 현재의 일부 LMM은 EU의 일반 데이터 보호 규정과 같은 주요 데이터 보호법을 위반할 가능성이 있다. 이러한 권리/보호/요구사항은 AI 개발의 가이던스가 되어야 한다.

LMM의 위반 사항은 다음과 같다: (1)개인 데이터의 정당한 사유 없는 무단 수집 및 사용, (2)데이터 사용 인지, 실수 수정, "잊힐 권리", 또는 데이터 사용 반대 권리를 보장하지 않음, (3)민감한 데이터 사용에 대한 투명성을 충분히 확보하지 못함. 법적으로는 사용자가 채팅 기록 데이터를 삭제할 수 있어야 함, (4)미성년자에 대한 “연령 게이팅”이 없음, (5)개인정보 유출, (6)부분적인 환각으로 인한 부정확한 개인 정보 게시. 데이터 보호 규정의 또 다른 위반 사례는 "설명 요구 권리"와 관련이 있다. 이는 개인 데이터를 자동화된 처리에 사용하는 시스템은 의사 결정 방식을 설명할 것을 요구한다. 일부 기업들은 “설명 가능성” 요구사항을 충족하도록 개발하고 있지만, LMM의 의사 결정 방식을 충분히 설명할 수 없다. 이외에도 심각한 위반 사항들이 있는데, 대부분 LMM의 학습 방법/사용 방식/데이터 관리 방식과 관련 있다. LMM이 데이터 보호규정이나 데이터 보호법을 완전히 준수할 수 없을 가능성도 존재한다.
많은 데이터 보호법 위반은 소비자 보호법 위반으로 간주될 수도 있다. 이러한 문제가 해결되지 않으면, 이는 보건 분야에서 AI 사용에 관한 WHO의 가이던스 원칙에도 직접적으로 위배될 수 있다. 이 가이던스에는 자율성 보호, 투명성 보장, “설명 가능성” 및 이해 가능성 원칙

이 포함된다. 기업들이 기존 법률을 준수하지 못하는 것은 AI 규제와 관련하여 일부 기업들이 심각한 우려를 표명하는 이유이다. 예) EU가 계획 중인 AI 규제 법안에 대응하여 한 기업대표는 해당 규정의 준수 어려움 때문에 유럽에서 자사의 핵심 LMM 제품을 제공할 수 없을 것이라고 밝혔다. 이러한 통첩은 개인정보 권리와 기타 보호 조치의 약화를 초래하거나, 일부 인권을 포기하려는 국가의 의지에 따라 의료 서비스 제공이 달라질 위험이 있다.

3.3 사회적 우려와 위험

LMM은 보건 시스템을 넘어 사회 전체적으로 영향을 미치기에 단일 법률/정책으로 우려를 해소할 수 없다. 이러한 사회적 영향은 LMM 상용화 선도 기업의 권력/권한을 강화할 것이다. 또한, LMM 시스템이 소비하는 탄소와 물은 환경/기후에 부정적인 영향을 미친다. LMM은 "문화가 이를 안전하게 흡수할 수 있는 속도보다 빠르게 수십억 명의 삶에 얽혀들며", 의료 및 의학 분야를 포함한 다양한 영역에서 사용되고 있다. 또한, LMM이 기술을 통한 젠더 기반 폭력을 증폭시킬 수 있다는 심각한 우려가 있다. 여기에는 사이버 괴롭힘/혐오 발언/"딥페이크" 등이 포함된다. 이러한 일은 소녀와 여성의 보건과 복지에 심각한 부정적 영향을 미친다.

대규모 기술 기업과 관련된 과제: LMM의 등장은 AI를 개발/배포하는 소수 대규모 기술 기업의 지배력을 강화했다. 더 정교한 LMM을 개발할 인적/재정적 자원, 전문 지식, 데이터, 컴퓨팅 능력을 갖춘 기업과 정부는 소수에 불과하다. LMM에 대한 컴퓨팅 능력과 투자가 증가하고 있으며, AI에 대한 수요가 증가함에 따라 "AI 인재"의 인건비도 높아지고 있다. LMM의 훈련, 배포 및 유지보수 비용이 계속 증가함에 따라, 이러한 기술이 많은 제품과 서비스(의료 포함)의 기본 구성 요소가 되는 기술을 소수 기업이 독점하는 "산업 지배" 위험이 있다. AI 연구 분야에서는 이미 대규모 기업들이 대학과 정부를 배제하고 있다는 증거가 존재한다.
현재 "전례없는" 수의 AI 박사 졸업생이 기업에서 일하는 것을 선택하고 있다. 예) 2004년: 약 20%의 졸업생이 산업계로 진출, 2020년: 거의 70%가 산업계에서 활동. AI를 전문으로 하는 대학 교수들도 대학을 떠나 산업계에서 종사하고 있으며, 이러한 이동은 2006년 이후 8배 증가했다. 산업계는 컴퓨팅 능력과 대규모 데이터 사용 측면에서도 정부와 학계를 압도하고 있다. 2021년에는 산업계 모델이 학계 모델보다 29배 더 커졌다. 또한, 정부의 예산은 산업계의 지출에 비해 현저히 뒤처지고 있다. "2021년, 미국 정부는 AI에 15억 달러를 할당했으며, EU 집행위원회는 약 12억 달러를 지출할 계획이었는데, 산업계는 전 세계적으로 3400억 달러 이상을 지출했으며, 이는 공공 투자에 비해 압도적으로 많은 금액이다".
"AI 투입"의 지배는 대규모 기업들이 AI의 결과물과 결과를 지배하게 되었음을 의미한다. 산업계가 개발한 가장 큰 AI 모델의 점유율은 2010년 11%에서 2021년 96%로 증가했으며, 2000년에서 2020년 사이 산업계 공동 저자가 포함된 연구 보고서의 비중은 16% 증가했다. 산업계의 지배는 AI의 응용 및 사용 뿐만 아니라 초기 단계 연구의 우선순위까지 정의하게 되었다. 산업계의 지배와 정부의 투자 부족은 공공 이익에 중요한 AI 기술, 특히 의료 및 의학 분야에서 대안을 찾기 어렵게 하고 있다. 한편, 제약 산업에서는 정부, 비영리 단체 및 자선 단체의 연구 개발 투자(특히 신약 개발 초기 단계와 특정 치료제의 후기 개발 단계)가 상당히 이루어지고 있다. 기업들은 점점 더 경제와 사회 부문(의료 포함)을 지탱하는 시스템 운영을 감독하게 될 것이며, 시민과 공무원이 자신의 삶을 관리할 능력에 대한 우려가 커져 간다.
대안과 규제가 없는 상황에서, 대규모 기업들의 내부 의사결정 방식과 사회/정부와의 관계는 더욱 중요해지고 있다. 기업들은 Frontier Model Forum이나 정부와의 파트너십 프로그램을 통해, 또는 미국/EU 정부와의 여러 약속을 통해 우려를 다양하게 해결할 수 있다. 하지만 기

업들이 윤리에 대한 기업적 책임을 다하지 않을 가능성도 있다. 예) AI 모델 설계/개발의 내부 윤리 원칙 준수 위한 관련 특정 개발 활동의 지연/중단을 줄이기 위해 윤리팀 축소/제거. 몇몇 대규모 기업들은 Frontier Model Forum을 통해 LMM을 포함한 “첨단 AI 모델의 책임감 있고 안전한 개발”을 보장하고, “첨단 모델의 책임 있는 개발 및 배포를 위한 모범 사례 식별”과 “정책 입안자, 학계, 시민 사회 및 기업들과 협력하여 신뢰와 안전 위험에 대한 지식을 공유”하겠다고 약속했다. 또한, 미국 정부와의 자발적 약속을 통해 기업들은 유해한 편향 및 차별을 피하고 개인정보를 보호하겠다고 서약했다. 그러나 자발적 약속이나 파트너십이 강력한 윤리적 책임을 대체할 수 있을지는 명확하지 않다. 예) 한 기업 윤리팀은 새로운 LMM의 출시를 중단할 것을 권고했으나, 이후 관련 문서를 수정하고 문서화된 위험을 축소시켰다.
대규모 기업들은 의료 제품/서비스 개발에 대한 경험이 없으며, 이 분야에 전문화되어 있지도 않아, 의료 시스템/제공자/환자의 요구사항에 민감하지 않을 수 있으며, 전통적인 의료 기업이나 공중 보건 제공자에게 익숙한 개인정보 보호나 품질 보증과 같은 문제를 다루지 못할 수 있다. 그러나 시간이 지남에 따라 이러한 민감성이 개선될 가능성은 있다.
LMM을 개발 중인 많은 기업들은 (1) LMM의 위험과 이점을 평가하기 위한 증거, 데이터, 성능 및 기타 정보와 (2) 모델 내 매개변수의 수를 요구하는 경우에도 이를 충분히 제공하지 않고 있다. 또한, 이러한 모델을 사용하여 자체 제품 및 서비스를 개발하는 기업들 역시 윤리적 문제와 위험을 평가하는 방식, 도입된 안전장치, 이러한 안전장치에 대한 LMM의 반응, 그리고 기술 사용이 제한되거나 중단되어야 할 시점에 대한 정보를 공개하지 않는다. Foundation Model Transparency Index는 100가지 지표를 기준으로 대규모 언어 모델 개발 기업 10곳을 평가한 결과, “주요 모델 개발 기업 중 어느 곳도 적절한 투명성을 제공하지 못하고 있으며, AI 산업에서 근본적인 투명성 부족을 드러내고 있다”고 지적했다. 미국 정부와 여러 대규모 기업 간의 자발적 합의에는 투명성과 관련된 두 가지 약속이 포함되어 있다: (1) 산업, 정부, 시민 사회, 학계와 위험 관리 정보를 공유하고, (2) AI 시스템의 역량, 한계, 적절하거나 부적절한 사용 영역에 대해 공개적으로 보고하기로 약속한다. 이러한 약속은 현 상황을 개선할 가능성이 있지만, 자발적이며 각 기업의 해석에 따라 달라질 수 있기 때문에 구체적인 규제 요구사항 없이는 충분한 공개로 이어지지 않을 가능성이 있다.
기업들은 내부 상업적 압박이나 외부 경쟁으로 인해 LMM이 어떻게 작동하는지 완전히 이해하기 전에, 그리고 적절한 테스트, 안전장치, 윤리적 위험 및 우려를 식별하고 해결했는지 여부에 관계없이 최대한 빠르게 시장에 출시하려 하고 있다. 기업들은 인터넷 검색과 같은 특정 분야에서 LMM의 시장 점유율이 수익을 창출할 수 있기 때문에 “선점자 우위”를 추구한다. 예) 검색 엔진에서의 시장 점유율 1%는 추가로 20억 달러의 수익을 의미한다. 또 다른 주요 기업의 임원은 자사의 LMM이 "완벽하지 않다”고 인정하면서도 “시장 수요가 있기 때문에 출시할 것”이라고 밝혔다. 위험을 완전히 식별, 검증, 평가, 완화하지 않고 LMM을 출시하는 기업들은 “윤리적 부채”를 축적하게 되며, 이러한 기술의 부정적 영향에 가장 취약한 기업들이 그 결과를 감당해야 한다. Frontier Model Forum의 회원들은 “AI 안전 연구를 증진”하고 “모범 사례를 식별”하는 데 헌신할 것을 약속했으며, 미국 정부와의 자발적 약속에는 출시 전에 AI 시스템에 대한 내부 및 외부 테스트가 포함된다.
상업적 압박은 기업이 LMM을 가능한 한 빨리 시장에 출시하게 만들 뿐만 아니라, 수익을 창출할 수 있는 서비스를 우선시하기 위해 공중 보건에 상당한 이익을 제공하는 제품 및 서비스를 우선시하지 않게 한다. 2023년, 한 기업은 단일 단백질 서열로부터 단백질 구조를 예측하는 단백질 언어 모델 ESMFold를 개발한 팀을 “해체”했다. 이 모델은 6억 개 이상의 단백질

구조를 포함한 데이터베이스를 생성했다.-운영 비용 혹은 서비스 유지 의지 문제?-

LMM의 탄소 및 물 발자국: LMM은 대량의 데이터를 학습시키는 데 상당한 에너지를 소비한다.-새 LMM을 훈련하는 데 약 2개월 동안 3.4 GWh의 에너지가 사용되었는데, 이는 미국 가정 300가구의 연간 에너지 소비량이다.- 대부분의 AI 모델은 화석 연료로 전력을 공급받는 전력망에서 훈련된다. 더 많은 기업들이 LMM을 도입하면 전력 소비는 계속 증가할 것이며, 이는 결국 기후 변화 위기로 다가온다. WHO는 기후 변화를 세계적 보건 문제로 간주하며, 이의 우선적 해결을 강조한다. 2030년에서 2050년 사이, 기후 변화로 인해 영양실조/말라리아/설사병/열 스트레스로 인해 매년 약 25만 명이 사망할 것으로 예상된다. 2030년까지 보건에 대한 직접적인 피해 비용은 매년 20억~40억 달러로 추정된다. 저/중소득 국가의 보건 인프라가 취약한 지역은 준비와 대응을 위한 지원 없이 이러한 상황에 대처하기 어렵다.
LMM은 상당한 물 사용 발자국을 가지고 있다. 한 기업에서 초기 LMM을 훈련시키는 데 70만 리터의 물이 소비되었다고 한다. 많은 개발자들이 탄소 발자국에 대해서는 점점 더 인식하고 있지만, 물 발자국에 대해서는 여전히 인식이 부족한 경우가 많다. 예)) LMM과의 짧은 대화(20~50건의 질문과 답변)는 약 500mL의 물을 소비한다. LMM을 훈련시키는 데 드는 전체 물 발자국은 AI 서버 제조/운송/칩 제작을 포함한 모든 물 소비를 포함하면 훨씬 더 클 것이다. 데이터 센터는 지역 물 공급에 스트레스를 줄 수 있다. 예) 미국 오리건 주의 한 도시에서 데이터 센터가 해당 도시 전체 물 사용량의 25% 이상을 소비했다. 물 발자국을 추적하는 것은 어려운데, 이는 탄소 발자국에 대한 인식/측정/투명성이 높은 반면, 기업들이 물 발자국에 대해서는 동일한 수준의 투명성을 제공하지 않고 이를 측정하지 않는 경우가 많기 때문이다.

"위험한" 알고리즘-인간의 인식적 권위의 대체: LMM이 점점 더 지식의 원천으로 간주되는 그럴듯한 응답을 제공함으로써, 의료/과학/의학 분야를 포함한 인간의 인식적 권위를 약화시킬 수 있다. 그러나 LMM은 실제 지식을 생성하지 않으며, 자신이 "말하고 있는" 내용을 이해하거나 질문에 대한 도덕적 또는 맥락적 추론을 할 수 없다. 이러한 우려가 지속된다면, 사회는 컴퓨터 기반 추론의 결과에 대비하지 못할 수 있다. 기업들은 LMM의 위험성에 대해 경고하면서도, 안전장치/규제 감독 없이 이를 사회에 출시한다. 이는 인간이 지식 생산의 통제 능력을 약화시킬 뿐만 아니라, 의료/의학 등 인간이 지식을 안전하게 활용해야 하는 사회적 시스템에서 그 능력을 저하시킬 위험이 있다. 특히 자원이 부족한 환경의 사람들은 AI 시스템 훈련에 그들의 데이터가 사용되지 않았을 가능성이 높아 정확성이 떨어지는 응답을 얻게 된다. 이러한 그룹은 보건의료 전문가/의료 제공자가 없는 상황에서, 잘못된 또는 부정확한 LMM 응답을 맥락화하거나 수정할 수 없기에 AI 시스템의 조언을 따를 가능성이 높다.
LMM이 불완전/잘못된 정보를 공공 영역과 지식 기반에 지속적으로 배출할 경우, 결국 “모델 붕괴”로 이어질 수 있다. 이는 LMM이 부정확하거나 거짓된 정보로 훈련되어 공공 정보 출처를 오염시키는 현상을 말한다. LMM이 보건 의료/사회적 중요 영역에서 이점을 최대화하면서 이러한 시나리오를 방지하려면, 정부/시민 사회/민간이 이 기술을 공익으로 이끌어야 한다.

II 보건의료 및 의학 분야에서의 LMM 윤리 및 거버넌스

WHO 전문가 그룹이 정의한 윤리 원칙은 보건의료/의학 분야에서 LMM의 개발, 배포 및 사용 평가에 필요한 기본 윤리적 요구사항에 대한 가이던스를 제공한다. 이는 정부/공공기관/연구자/기업 및 실행자가 LMM의 사용을 관리하는 방식의 기반이 되어야 한다.
거버넌스는 정부와 국제 보건 기구 등 의사 결정자들이 보편적 보건을 보장하기 위해 방향을

설정하고 규칙을 제정하는 기능을 의미하며, 다양한 요구와 영향력을 조율하는 정치적 과정이기도 하다. 기존 법과 정책은 초기 LMM 개발 이전에 만들어져 이를 효과적으로 관리하기에 한계가 있다. LMM 거버넌스는 AI 거버넌스와 유사하게 법률, 윤리 원칙, 인권 의무, 실천 강령, 그리고 기업·산업·표준 기관의 내부 절차 등을 포함한다.

현재 LMM은 충분한 이해 없이 빠르게 확산되고 있으며, 초기 대응으로 개발 중단이 제안되기도 했다. 일부 국가는 사용을 제한하지만, 대부분은 적절한 거버넌스를 통해 사회적 이익을 추구하고 있다. 주요 AI 기업들도 신중한 개발을 촉구하지만, 정부와 기업 모두 경쟁과 상업적 압력에서 자유롭지 않다. 낙관론자들은 더 큰 데이터와 강력한 알고리즘이 문제를 해결할 수 있다고 주장하는 반면에, 비관론자들은 LMM의 한계가 구조적이며 훈련 데이터 및 모델 매개변수의 확대가 위험을 키울 수 있다고 본다.

LMM 거버넌스는 급속한 발전과 증가하는 용도에 맞추어야 하며, 기술적 우위 추구 정부나 상업적 이익 추구 기업이 특권을 가지면 안된다. 초기 권고사항은 윤리 원칙과 인권 의무를 거버넌스의 중심에 두고 있으며, 이는 기업의 절차/관행과 정부 제정 법률/정책을 포함한다.

LMM은 여러 행위자의 연속적인 결정으로 만들어지며, AI 가치사슬의 각 단계에서 내린 결정은 개발, 배포, 사용에 영향을 미친다. 이러한 결정은 국가·지역·글로벌 차원의 법과 정책에 의해 규제될 수 있다. AI 가치사슬은 데이터, 연산력, 전문 지식 등의 인프라를 바탕으로 범용 모델을 개발하는 데서 시작되며, 이 모델은 의료 등 다양한 작업에 활용될 수 있고, 일부는 의료 분야에 특화되어 훈련되기도 한다.

의료 및 의학 분야에서 사용되는 LMM의 적절한 거버넌스는 데이터 수집부터 의료 응용 프로그램 배포까지 AI 가치사슬의 세 가지 중요 단계와 질문:

•(설계 및 개발 단계) 범용 기반 모델의 설계 및 개발

-관련 위험을 가장 잘 해결할 수 있는 행위자는 누구이고 해결해야 할 위험은 무엇인가?

•(제공 단계) 범용 기반 모델을 활용한 서비스, 응용 프로그램 또는 제품의 정의

-관련 행위자들이 이러한 위험을 어떤 윤리 원칙을 준수하면서 어떻게 해결할 수 있는가?

•(배포 단계) 의료 서비스 응용 프로그램 또는 서비스의 배포

-정부는 어떤 법률, 정책 또는 투자를 통하여 어떤 역할로 이러한 위험을 해결할 수 있는가?

설계/개발 단계는 개발자의 윤리적 약속/규범 준수 관행과 정부 정책 및 투자가 중요하고, 제공 단계는 의료 및 의학 분야에서 LMM의 사용을 평가하고 규제하기 위한 정부의 조치가 중요하다. 배포 단계에서는 사용자의 피해를 식별하고 방지하기 위한 조치가 있어야 한다.

4 범용 기반 모델(LMM)의 설계 및 개발

범용 기반 모델은 방대한 데이터와 컴퓨팅 능력, 과학/공학분야 인적 자원도 필요하다. WHO 의료 분야 AI 윤리 및 거버넌스에 관한 가이던스는 의료용 AI 개발자가 “제품의 설계, 감독, 신뢰성 및 자율 규제를 개선하기 위한 조치에 투자해야 한다”고 권고하고 있다. 다음 대부분의 연구 결과와 권고사항은 범용 기반 모델에 적용되지만, 이 가이던스는 특히 의료 분야에서 사용될 모델에 초점을 맞추고, 의료 LMM의 설계와 사용을 안내하는 데 목적이 있다.

4.1 범용 기반 모델(LMM) 개발 시 해결해야 할 위험

범용 기반 모델의 설계와 개발에서 발생할 수 있는 위험은 사회적으로 영향을 미치고 부정적 결과를 초래할 수 있으며, 이를 해결하는 책임은 개발자에게 있다. LMM의 훈련 데이터 결정, 데이터 보호, 편향 완화 등의 결정은 개발자만이 내릴 수 있으며, 이후 제공자나 사용자에 의

해 완화되지 않는다-예) LMM 훈련 데이터와 관련된 결정, 데이터 보호와 품질 보장 의무, 편향 완화 조치, “AI로 인한 유해성” 생성 방지-. 이러한 설계 결함에 대한 책임을 개발자에게 묻지 않으면 자원을 가진 기업들이 문제를 무심코 내재화할 수 있다.

개발자의 최소 여덟 가지 위험 (정부의 법률 및 규정 제정으로 해결 가능): 편향(설계와 훈련 데이터와 관련됨)/개인정보(훈련 데이터 및 기타 입력 데이터의 보호)/노동 문제(불쾌한 콘텐츠를 제거하기 위한 외주 데이터 필터링)/탄소 및 물 발자국/허위 정보, 혐오 발언 또는 잘못된 정보/안전 및 사이버 보안/인간의 인식적 권위 보존/LMM의 독점적 통제

4.2 개발자가 범용 기반 모델(LMM)과 관련된 위험을 해결하기 위해 취할 수 있는 조치
개발자는 윤리 약속이나 정부 요구 사항에 대한 다양한 조치를 통해 위험을 해결할 수 있다.

AI 전문성(과학 및 공학): 개발자는 과학 및 SW 인력의 위험 식별 및 회피를 보장해야 한다. WHO 윤리 가이던스는 인력의 교육과 설계 과정의 포괄성에 대한 권장 사항을 제시했으며, WHO 전문가 그룹은 “고위험 AI 개발자 대상 라이선스 또는 인증”을 권고했다.

의료 분야에 사용될 수 있는 LMM을 개발하는 기관은 신뢰 확보를 위해 인증이나 교육을 고려해야 하며, 관련 표준은 정부와 협력해 WHO 윤리 원칙에 부합하게 마련돼야 한다. 또한, 의도하지 않았더라도 의료 사용이 예측된다면 개발자는 이에 대비한 전문성을 갖춰야 한다.

데이터: 데이터는 가장 중요한 인프라 요구사항이다. LMM 훈련 데이터의 품질과 유형은 해당 모델의 핵심 윤리 원칙과 법적 요구 사항 충족 여부를 결정한다. 고품질 데이터의 중요성과 데이터 작업이 종종 과소평가 되며, 이는 의료 및 의학과 같은 “고위험” 분야에서 AI에 중대한 부정적 영향을 미칠 수 있다. 저품질 혹은 편향 데이터는 인간의 복지, 안전, 공익 증진, 포괄성과 형평성 보장이라는 WHO의 여러 가이던스 원칙을 위반할 수 있다.

의료 데이터를 사용하는 LMM 개발자는 사전 동의의 엄격한 법적 요구로 인해 작은 데이터 세트를 활용할 수 있다. 이는 품질과 다양성을 확보하고 서비스 대상을 반영한 데이터 구성에 유리하지만 재식별 위험을 높일 수 있다. 또한, 소규모 데이터는 환경 부담을 줄이고 소규모 단체의 참여를 촉진하는 이점도 있다. 데이터 크기와 무관하게, 개발자는 처리 전 개인정보 보호 영향 평가를 수행해야 한다. 그리고, 저/중소득 국가에서의 데이터 수집은 “데이터 식민주의”로 간주될 수 있다.

평가가 개인정보 보호와 데이터의 편향/정확성 등 품질에 대하여 이루어져야 하지만, 연구자들은 데이터 감사의 어려움과 고비용 때문에 투자를 꺼린다. 개발자는 고품질 데이터와 데이터 보호법 준수를 위해 신중한 데이터 수집이 필요하며, 데이터 브로커 등 제3자 출처는 피해야 한다. 부적절한 데이터는 편향/오류/법적 위반으로 이어져 LMM이 불법으로 간주될 수 있다. 제3자 데이터 사용 시 신뢰성/정당성을 위한 인증이 필요하다. 모든 학습 데이터는 최신 상태를 유지해야 하며, 특히 의료 분야에서는 최신성과 정확성이 모델 성능에 매우 중요하다. 데이터의 투명성을 충분히 확보하기는 어려워지고 있으며, 주요 AI 기업들은 경쟁과 안전 문제를 이유로 모델의 세부 정보를 공개하지 않고 있다. 데이터 비공개는 WHO의 투명성과 설명 가능성 원칙에 어긋난다. 개발자는 데이터를 투명하게 하여 LMM 미세 조정 혹은 의료 응용 프로그램을 개발하거나 직접 사용하는 자들에게 데이터의 한계를 인식시켜 주어야 한다.

개발자는 저소득 국가에서 데이터 작업 시, 생활임금 지급, 정신 건강 지원, 안전장치 마련 등 작업자 보호 조치를 취해야 한다. 정부는 이를 위해 노동 기준을 강화하고 공정한 경쟁 환경을 조성하며 시간이 지남에 따라 노동 기준이 유지되고 개선되도록 보장해야 한다.

윤리적 설계 및 가치 기반 설계: AI 기술 개발에 윤리 및 인권 기준을 통합하도록 "가치 기반 설계"로 인간의 존엄성, 자유, 평등, 연대와 같은 가치를 설계의 기반으로 삼고, 이를 비기능적 요구사항으로 간주한다. WHO의 전문가 가이던스에 제시된 AI 기술 설계에 대한 여러 권장 사항 중 "가치를 위한 설계"를 포함한 일부는 다시 강조할 필요가 있다.

이 가이던스는 LMM 개발에 과학자, 잠재적 최종 사용자와 모든 직간접적 이해관계자가 AI 개발 초기 단계부터 포괄적이고 투명한 설계에 참여하고 윤리적 문제 제기, 우려 표명, 검토 AI 응용 프로그램에 대한 의견 제시의 기회를 가져야 한다"고 권장했다. 예를 들어, 환자 대표를 포함한 "인간 감독 대학" 도입이 제안되며, 이는 자동화 편향을 줄이고 WHO의 포괄성/형평성 원칙을 강화할 수 있다. 포괄적 설계는 연령/능력/인종/민족/성별 또는 성 정체성 같은 다양한 관점을 포함하여 포괄성/형평성을 보장하는 WHO 원칙을 촉진한다.

WHO 초기 가이던스는 또한 "설계자와 기타 이해관계자들은 보건 시스템의 역량 개선과 환자 이익 증진을 위해 필요한 정확성과 신뢰성을 지니도록 잘 정의된 작업을 AI 시스템이 수행하게 설계하여야 한다"고 권장했다. 이를 위해 개발 전 "사전부검"으로 "가상 실폐"를 고려한 역설계와 "레드 티밍(red teaming)" 평가에 의한 취약점 식별을 하도록 한다. 이러한 시뮬레이션은 위험을 사전에 파악하고 모델의 안전성과 신뢰성을 강화하는데 도움이 된다. 또한, WHO 초기 가이던스는 "디자이너가 '가치 기반 설계'를 위해 사용하는 절차는 합의된 원칙, 모범 관행, 설계 윤리 기준 및 발전하는 전문적 규범에 의해 정보를 제공받고 업데이트되어야 한다"고 권장했다. 이는 개인정보 보호, 환경 영향(탄소·물 발자국) 완화, AI 생성 콘텐츠 표시 등을 통해 인간 중심의 권위를 유지하고 인식적 권위의 중심을 대체되지 않도록 한다. 즉, LMM이 유용한 정보를 생성할 수는 있지만, 지식 생산을 대체할 수는 없음을 상기시킨다.

환경적 우려 고려 설계: 개발자는 모델의 에너지 효율성을 개선하는 등 에너지 소비를 줄이기 위해 가능한 모든 조치를 취해야 한다. 여러 대규모 기업이 이러한 방식을 추구한다.

에너지 효율성을 개선하는 또 다른 접근 방법으로, 작은 데이터 세트로 훈련된 소형 LMM은 에너지 효율성이 높고, 비용이 적게 들어 소규모 단체도 접근 가능하다. 특히 의료나 과학 등 특정 목적의 전문화된 LMM 개발에 유용하며, 정확도 향상에도 기여할 수 있다.

4.3 정부 법률, 정책 및 공공부문 투자

범용 기반 모델의 위험을 줄이고 방지하기 위해 법률·정책이 시행되며, 정부는 윤리적 설계·개발의 촉진 및 지원을 위해 공공 투자를 할 수 있다.

데이터 사용 규제 법률 및 정책: 데이터 보호법은 일반적으로 권리 기반 접근방식을 기반으로 하며, 데이터 처리 규제 기준을 포함하여 개인의 권리를 보호하고, 공공 및 민간 데이터 관리 책임자와 처리자의 의무를 설정하며, 법적 권리를 침해하는 행위에 대한 제재와 구제 조치를 지녀야 한다. 이 법은 150개국 이상이 도입했으며, AI 기술 개발의 기반이 되고 있으나, 대부분 생성형 AI 등장 이전 제정으로 적용 한계 때문에 엄격한 적용은 피하려 한다.

LMM 훈련에 사용하는 보건 데이터를 포함한 데이터는 적법하게 수집·처리돼야 하며, 이를 위해 당사자의 명확한 동의가 필요하고, 추가 활용 시에는 별도의 법적 근거가 요구된다. 일부 기업은 동의 없이 수집된 데이터의 사용 혐의로 조사받고 있고, 방대한 데이터 수요는 개발자가 법적 요건을 무시하게 할 수 있다. 이는 WHO의 인간 자율성 원칙에 어긋난다. WHO 전문가 그룹은 정부에 "보건 데이터 사용 및 개인 권리 보호, 특히 의미 있는 동의를 받을 권리를 포함한 명확한 데이터 보호법과 규정을 마련해야 한다"고 권장한다. 중국은 2023년 8월 생성형 AI 규정을 시행해 LMM 훈련 데이터의 수집과 사용을 감독하고 있다. 규정에 따르면, 제공자는 편향 방지/명확한 라벨링/데이터 품질 확보를 위한 조치를 해야 하며, 개발자는 데이터의 진정성·정확성·다양성 확보에 노력해야 한다. 이 조치는 중국 내 서비스 제공 기업에만 적용되는데, 기업에는 일정 수준의 유연성이 주어진다.

데이터 관련 법적 조항은 기반 모델 훈련 데이터 소스와 데이터 거버넌스에 대한 요구사항이 포함될 수 있다. 정부의 설계/개발 단계 추가적인 조치는 다음과 같다.
•대상 제품 프로필: 정부와 국제기구는 의료 분야 사용 LMM의 특성과 기준을 명시한 대상 제품 프로필을 제시할 수 있으며, 특히 공공 보건용 기술 구매를 염두에 둘 때 유용하다.
•설계 및 개발 표준과 요구사항: 정부는 범용 기반 모델의 전 생애 주기에 예측 가능성/해석 가능성/수정 가능성/안전성/사이버 보안 등 특정 결과를 달성하도록 요구할 수 있다.
•사전 인증 프로그램: 규제 기관은 사전 인증 프로그램 등으로 개발자에게 법적 의무를 부여하고, 편향이나 자율성 침해 같은 윤리적 위험을 식별·회피하도록 인센티브를 제공할 수 있다. WHO는 가이던스로 사전 인증을 권장한다.
•감사(Audits): 정부는 LMM 개발 초기 단계의 감사(거버넌스 감사, LMM 자체 감사, 응용 프로그램 감사 등)를 도입할 수 있으며, 이는 의료 분야에서 사용될 LMM 승인 요구사항에 통합될 수 있다. 효과적 감사를 위해서는 품질이 의도한 목적을 충족하는지 평가해야 한다.
•환경 발자국(Environmental Footprint): 정부는 개발자에게 에너지 소비 측정, 에너지 사용 절감, 환경 기준 충족 등 탄소 및 물 발자국 문제를 해결하도록 요구할 수 있다.
•"기계 생성" 콘텐츠 공지: 콘텐츠가 기계 생성임을 사용자에게 공지토록 요구할 수 있다.
•정부 조치: 정부는 의료용 AI의 조기 등록을 요구하거나 인센티브를 통해 장려할 수 있다. 이는 부정적 결과 공개, 출판 편향/낙관적 해석 방지, 유익한 지식 통합에 도움이 된다.

공익을 위한 LMM 개발을 지원하는 공공 인프라: 의료 분야에서 LMM 활용이 늘면서, 정부는 윤리 원칙을 준수한 개발을 촉진하기 위해 컴퓨팅 자원과 공공 데이터를 포함한 공공 인프라를 제공할 수 있다. 이 인프라는 다양한 개발자가 접근 가능하며, 윤리 기준 준수를 조건으로 할 수 있다. 이는 독점 방지와 공정한 경쟁 환경 조성에도 기여한다.

정부는 독립적으로 의료용 LMM 개발을 위한 인프라를 구축할 수 있다. 프랑스 정부의 지원과 학계·기업 협력으로 1,750억 매개변수의 BLOOM 모델이 약 700만 달러의 비용으로 개발되었다. 공정한 경쟁을 위해 학계 지원도 필요하다. 캐나다: 국가고급연구컴퓨팅 플랫폼을 통해 학계를 지원, 중국: 공공 데이터와 컴퓨팅 접근을 위한 네트워크를 구축, 미국: 공공 연구 클라우드와 데이터 세트를 제안, 유럽 시민사회: 정부의 AI 인프라 및 연구 지원 확대를 촉구.

4.4 오픈소스 LMM

오픈소스 LMM이 윤리적 원칙과 위험 해결에 얼마나 기여할 지는 불확실하지만, 일반적으로 투명성과 참여를 높이는 데 도움이 된다. 소스 코드 공개를 통해 사용자는 시스템을 이해하고

문제를 식별하며 개선할 수 있다. 오픈소스는 독점적/폐쇄적이지 않아 비영리 단체 등도 저비용으로 접근이 가능하며, 공개된 코드와 데이터는 검토와 감시를 통해 모델의 신뢰성과 견고성을 높일 수 있다.

그러나 오픈소스 LMM의 지속 가능성은 불확실하다. 많은 개발이 Meta의 제한적 공개 모델에 기반했지만, 기업들이 공개를 중단하면 오픈소스 생태계도 위축될 수 있다. Meta는 개방성이 혁신과 시장 발전에 기여한다고 강조했으나, 독립적인 관찰자들은 Meta가 비상업적 사용에 제한을 두어 진정한 오픈 소스와는 거리가 있다고 지적했다. 개발자에게 성능 모니터링의 부담이 있고, 오픈소스 모델의 이점이 규제와 보안 문제를 완전히 대체할 수 없으며, 오용과 공격에 취약하다. 최근 연구에 따르면 오픈소스뿐 아니라 폐쇄형 시스템의 안전장치도 우회 가능한 것으로 나타났다. 결국, 오픈소스 모델도 여전히 블랙박스 기술에 기반하고 있다.

오픈소스 LMM을 장려하는 한 가지 방법은 정부가 공공 자금이나 지식재산으로 개발된 LMM을 널리 공개하도록 요구하는 것이며, 이는 기존 공공 연구의 공개 접근 방식과 유사하다. 또한, 공공 시설과 감독 하에 오픈소스 R&D를 장려함으로써 무분별한 유출보다 더 책임 있는 대안을 제시할 수 있다.

5 범용 기반 모델(LMM)을 활용한 서비스 제공

범용 기반 모델의 의료 사용은 사용자가 결과물을 요청하거나 개발자가 이를 응용 프로그램에 통합하는 방식에 따라 달라지며, 두 경우 모두 새로운 위험이 발생한다. 이는 개발자와 제공자에 의해 해결되어야 하며, 정부는 배포 전 이 기술의 사용을 평가하고 규제할 책임이 있다.

5.1 범용 기반 모델을 사용하는 보건의료 서비스 및 응용 프로그램 제공 시 해결해야 할 위험

범용 기반 모델의 의료 사용을 평가·승인해야 하는지는 의견 차이가 존재한다. 일부 기술 기업은 정부가 모델 자체보다는 고위험 응용 프로그램에 집중하도록 비공식적으로 설득해 왔다. 이는 범용 기반 모델 제공자와 사용자 모두와 관련 있다. 기업들은 모델에 대한 규제가 개발자에게 과도한 책임을 지우며, 가치사슬 내 다른 주체들도 책임을 져야 한다고 주장한다.

범용 기반 모델 개발자에게 모든 책임을 지우는 것은 부적절할 수 있지만, 책임을 제공자나 사용자에게만 전가하는 것도 문제다. 이들은 모델 개발에 관여하지 않았고, 한계나 위험을 잘 모를 수 있으며, 막강한 권한과 자원을 가진 개발자들이 책임을 회피하게 되어 의료 AI 규제에 큰 허점을 남길 수 있다. 개발자는 LMM을 의료 목적(또는 다른 용도)로 사용하지 않으려고 할 수도 있다. 이 경우, 보건용 응용 프로그램에서의 사용을 제한하거나, 의료 목적의 사용 시 경고 메시지를 제공하고 적절한 정보나 서비스로 사용자를 안내할 수 있다. 이런 조치가 없거나 개발자가 의료 분야에 LMM을 적용하려 한다면, 개발자는 고유한 책임을 지게 되며, 개발자와 제공자 모두 관련 위험을 해결할 추가 의무를 갖게 된다. 이런 책임은 AI 시스템의 의료 분야 사용 여부를 결정하는 정부의 법과 정책에 따라 정의된다. 개발자와 공급자는 의료 분야에서 LMM을 사용하려면 공동 책임을 져야 하며, 법이 없거나 미비한 경우 계약으로 책임이 정해질 수도 있다. 배포 전에 해결해야 할 주요 위험으로는 편향, 잘못된 정보 또는 환각, 개인정보 보호, 조작 위험, 자동화 편향 등이 있다.

5.2 정부가 도입할 수 있는 위험 해결 조치 및 준수해야 할 윤리 원칙

LMM과 응용 프로그램이 빠르게 개발되고 있기 때문에, 정부는 보건의료 시스템이나 과학·의

학적 목적에 AI를 활용하기 위한 규제/기준을 신속히 마련해야 한다. 이는 의료기기나 의약품을 다루는 기존의 규제 기관이 AI 기술을 평가하고 승인하거나, 새로운 기관을 설립하는 방식이 될 수 있다. 다만 저/중소득 국가는 자원 부족과 기존 기관의 과부하로 어려움이 있다.

일부 고소득 국가는 기업들과 협력해 자국 기반 모델의 자발적 평가/결과 공개를 하고 있다. 이를 통해 대중과 연구자에게 정보를 제공하고, 기업이 오류를 수정토록 유도하지만, 이런 자발적 접근은 한계가 있고 지속 가능하지 않을 수 있다. LMM과 관련된 평가는 의료 시스템에서 쓰이는 AI뿐만 아니라, 임상과 웰니스 사이의 애매한 영역에서 사용되는 LMM과 응용 프로그램의 위험도 함께 다뤄야 한다. 기술이 빠르게 확산되고 있기 때문에, 정부는 초기 단계에서부터 이런 응용 프로그램을 식별하고, 공통된 기준과 규제를 마련해 기준을 충족하지 않을 시 대중에게 배포되지 않도록 해야 한다. 또한 개발자와 제공자는 보건의료용 AI 기술이 법과 정책의 최소 요건을 충족함을 입증해야 한다. LMM 관련 위험과 과제를 고려할 때, 이 기술이 안전하고 효과적일 것이라거나 기존 접근방식보다 우수할 것이라고 가정하면 안 된다.

다음은 의료용 LMM 사용에 적용될 수 있는 여러 법률, 정책 및 요구 사항이다.

공개(투명성) 요구 사항: 적절한 규제를 위해서 정부가 평가/승인 능력을 갖추는 것뿐만 아니라, 평가에 필요한 정보도 충분히 확보해야 한다. 이는 AI 기술을 안전하게 관리하고, 가치사슬 내 다른 주체들이 기술을 책임 있게 사용할 수 있도록 한다. 예) 개발자가 모델의 성능이나 환각 경향성을 공개하지 않으면, 제공자는 모델을 적절히 조정하거나 신중한 마케팅에 필요한 정보를 얻지 못할 수 있다. 이 정보 공개는 의료 사용자들이 LMM 사용 시 부정확한 정보를 피하거나 결과를 더 신중히 검토하는 데 도움을 줄 수 있다.

공개와 투명성은 WHO의 핵심 원칙이며, AI 시스템의 설명 가능성과 이해력 증진에 중요하고, 범용 기반 모델과 응용 프로그램 평가 시, 이 요소가 요구되어야 한다. WHO는 정부가 AI 기술의 안전성/효과성 감독을 위한 소스 코드/데이터 입력/분석 방법 등 특정 정보의 투명성을 권고한다. LMM 관련 공개 항목에는 내부 테스트 결과/탄소·물 발자국/학습 결과물, 그리고 LMM이 학습 시 얻은 지식의 이해도를 높이는 '개방형 가중치' 기준 등이 있다. 여러 형태의 공개는 제공자, 사용자, 규제 기관에 유용하기에 명확히 표시되어야 한다. 예) 기반 모델의 능력과 한계를 설명, 벤치마크에 따른 평가, 내부·외부 테스트 결과 보고.

데이터 보호법: LMM 개발과 학습 데이터 관리 방식이 데이터 보호법을 위반할 가능성이 있다. 또한, LMM이나 응용 프로그램에서 민감한 개인 정보가 실수로 공개될 수 있다.

데이터 공개는 개발자의 자율성 보호 책임의 위반에 해당하며, 민감한 데이터를 허용된 기간 이상 보유하면 데이터 보호 법률 위반이 될 수 있다. 일부 개발자는 사용자가 챗봇 성능 향상을 위해 제공한 콘텐츠를 제외할 수 있는 선택권을 제공한다. 정부는 LMM 입력 데이터를 보호하기 위한 규칙을 설정하고 시행해야 한다.

보건의료 사용 범용 기반 모델 및 응용 프로그램 평가-인권법 대 위험 기반 프레임워크: AI 기술의 평가/규제를 위한 법적 프레임워크가 마련되고 있으며, AI가 인권(EU의 '기본권')을 충족해야 하는지 혹은 위험 기반으로 평가할지에 대한 논의가 있다. EU는 AI 법에서 위험 기반 접근을 채택했으며, 기술의 위험 수준에 따라 요구사항과 입증책임이 달리 부과된다.

AI가 의료에 사용될 경우, 인권과 윤리 기준은 반드시 지켜져야 하며, 이는 범용 기반 모델에도 적용된다. 위험 수준과 관계없이 모든 기술은 검토 대상이며, 개발자와 제공자는 이를 준수해야 하고, 필요한 경우 인권 영향 평가도 실시해야 한다. WHO는 정부가 AI 시스템 전 생애 주기에 걸쳐 윤리, 인권, 안전, 데이터 보호를 다룬 영향 평가를 의무화해야 한다고 권장한다. 이 평가는 AI 도입 전후 독립된 제 3자에게 받아야 하고, 결과는 민감 정보를 고려해 공개되어야 한다. 제 3자에 의한 평가 수행의 품질과 신뢰성도 철저히 검토되어야 한다. 영향 평가를 통해 AI 기술이 편향을 유발하거나 개인정보를 노출하거나 사용자를 조작할 가능성이 있는지 확인한다. 개인정보 보호는 개발자와 제공자가 함께 해결해야 한다. 또, 영향 평가는 자동화된 의사결정을 피하고 인간의 개입을 보장해, LMM의 출력에 대한 과도한 의존이나 잘못된 정보를 받는 상황, 혹은 자동화 편향을 예방할 수 있다.

정부는 보건의료에서 LMM에 대해 위험 기반 프레임 워크를 채택할 수 있지만, 이 방식이 인권 중심 접근을 대체할 수는 없다. 정신건강 조언 같은 고위험 기능이나 취약 집단에 의해 사용되는 AI 기술은 높은 입증 책임을 진다. 저위험처럼 보이는 AI 기술도 결국 피해를 일으킬 수 있어 주의가 필요하다. AI 규제 평가를 모든 기반 모델에 적용할지, 아니면 주요 기반 모델에만 적용할지, 그리고 언제 제공자에게 적용할지를 놓고 추가 논의가 이어지고 있다.

본 가이던스는 모든 기반 모델에 대해 위험 기반 및 인권 기반 평가를 적용해야 하는지, 또는 시스템 기반 모델에만 적용해야 하는지에 대한 권장 사항을 포함하지 않는다. 전문가 그룹은 모든 LMM에 평가를 적용할 경우 대기업들이 유리해질 우려를 제기했으며, 이는 당국의 관심을 끌고 있다. 당국은 LMM 개발 기업들의 사용 관행에 주목하고 있으며, 향후 추가 조사가 예상된다. 제공자는 LMM의 목적과 기능을 변경할 수 있어 AI 규제 평가를 받아야 한다. 보건의료나 의학 용도로 범용 기반 모델을 사용하는 경우, 개발자와 제공자는 해당 요구사항을 준수해야 하며, 변경이 크거나 개발자의 통제를 벗어날 경우 제공자의 규제 책임이 더 커진다.

의료기기 규제: 정부는 범용 기반 모델이나 응용 프로그램의 의료기기 간주 여부를 결정할 수 있다. 일반적으로 개발자가 의료 목적을 주장하지 않으면 의료기기로 간주 되지 않지만, 의료 목적으로 개발되거나 수정된 LMM은 의료기기로 분류될 가능성이 높다,

LMM 기반 챗봇이 의료 상담을 제공하면, EU 및 미국 규제에 따라 의료기기로 분류될 가능성이 높다. WHO는 AI 시스템의 성능을 테스트할 때, 기존 데이터 세트와의 실험실 비교만 의존하지 말고 무작위 시험에서 건전한 증거를 요구해야 한다고 권고했다. LMM이나 응용 프로그램이 의료기기로 규제되려면, 개발자와 제공자가 법적 요구 사항을 충족하는 증거를 제시해야 하며, EU와 미국의 AI 의료기기 규정은 설명 가능성, 편향 통제, 투명성 등을 포함할 가능성이 높지만, 현재 LMM 기반 챗봇은 이러한 기준을 충족하기 어렵다. 임상 의사결정을 지원하는 LMM은 실험적으로 사용되고 있으며, 면책조항이 있지만 의료기기 법률의 적용을 피할 수 없다. 이러한 실험은 허가된 임상 시험 환경에서만 이루어져야 하며, 정부는 통제된 실험적 사용을 규제 샌드박스에서 검토할 수 있다. 샌드박스는 실제 임상 환경에서 보호 장치와 감독 하에 실험을 허용하지만, 이는 공식 규제를 받는 국가에서만 적용 가능하다.

소비자 보호법: 정부는 LMM과 응용 프로그램이 사용자와 환자에게 피해를 주지 않도록 소비자 보호법을 개발해야 한다. 이 법은 조작적 관행, 차별, 편향을 방지하고, 기술 개발자가 부정적 결과를 해결하도록 요구할 수 있다. 또한, LMM이 인간처럼 보이도록 오도하는 표현

(예: “내 생각엔”)을 제한하거나 금지할 수 있다.

6 범용 기반 모델(LMM)을 활용한 배포

LMM은 윤리적으로 설계되고 적절한 규제를 거쳐도 상용화 시 위험이 따를 수 있다. LMM의 배포 주체는 개발자, 제공자, 정부, 병원, 보건 의료 및 제약 기업 등 다양하다.

6.1 LMM 사용 보건의료 서비스 또는 응용 프로그램 배포 시의 해결 필요 위험

LMM 배포 시 위험은 예측 불가능한 응답, 의도치 않은 사용 방식, 그리고 시간에 따른 출력 변화에서 비롯될 수 있다. 배포 시 해결해야 할 주요 위험은 다음과 같다: 부정확하거나 잘못된 응답/편향/LMM에 입력되고 출력되는 데이터의 개인정보/LMM의 접근성 및 경제성/노동 및 고용에 미치는 영향/자동화 편향과 기술 저하/의료 제공자와 환자 간 상호작용의 수준

여기서는 LMM 사용 시 위험 완화 방법/정부 규제 역할/의료 종사자 교육과 역량 강화를 다룬다.

6.2 배포 중 개발자와 제공자의 지속적 책임

LMM의 승인 후에도 개발자와 제공자는 지속적인 책임을 지며, 이는 특정 위험이 배포 이후에만 해결될 수 있기 때문이다. 이 의무는 법이나 규정으로 요구될 수 있다.

첫째, 정부는 LMM이 널리 배포될 경우, 제 3자가 인권 및 데이터 보호 관련 사후 감사와 영향 평가를 실시하고 그 결과를 사용자 특성별로 공개하도록 해야 한다. 둘째, 정부는 LMM이 부정확하거나 유해한 정보를 생성했을 때, 이를 방지하지 않은 개발자나 제공자에게 책임을 물을 수 있다. 셋째, 개발자와 제공자는 LMM의 안전한 사용을 위해 충분한 기술 문서와 지속적인 운영 정보를 제공해야 한다.

6.3 배포자의 책임

배포자는 LMM이나 응용 프로그램 사용과 관련된 위험을 방지하거나 완화할 책임이 있다. 우선, 배포자는 개발자나 제공자로부터 받은 정보를 바탕으로 LMM이 부적절한 환경에서 사용되지 않도록 해야 하며, 경고 받았음에도 불구하고 이를 무시하고 배포할 경우 발생한 피해에 대한 책임을 져야 한다. 둘째, 배포자는 LMM 사용으로 인한 위험과 오류를 사용자에게 명확히 알릴 책임이 있으며, 필요한 경우 피해 예방을 위해 사용 중단이나 시장 철수 조치 책임도 질 수 있다. 셋째, 배포자는 LMM의 경제성과 접근성을 개선해 누구나 이용할 수 있도록 해야 하며, 다양한 언어와 문자로 학습·배포되도록 개발자 및 제공자와 협력해야 한다.

6.4 정부 프로그램 및 관행

LMM을 보건의료 시스템에 도입하는 것은 의료 종사자에게 막대한 적응 노력이 요구되지만, 개발자나 제공자는 이를 지원할 충분한 교육이나 전문 인력을 갖추지 못한 경우가 많다

정부는 환자 권리를 보호하고 LMM이 적절히 사용되도록, 의료 종사자와 환자를 ‘인간 감독 협회(human oversight colleges)’에 참여시킬 수 있다. 정부, 대학, 병원 등은 의료 종사자들이 LMM을 효과적으로 사용하고 적절히 훈련받도록 해야 한다. 교육 내용에는 (i) LMM의 결정 방식과 한계, (ii) 우려 사항 식별, (iii) 자동화 편향 회피, (iv) 환자와의 소통, (v) 사이버 보안 등이 포함된다. 의료 종사자 교육은 중요하며, LMM 조언을 의료 결정에 활용할 경우 환자나 일반인에게 그 사실과 관련 위험을 충분히 설명하고 사전 동의를 받아야 한다. 또한, 의

료 종사자 교육은 LMM 사용 시 법률 위반을 방지하는 데 중요하다. 예를 들어, 보호된 환자 정보를 LMM에 입력하면 HIPAA 등 법률을 위반할 수 있으며, LMM이 신뢰받는 도구로 인식될수록 무의식적으로 환자 정보를 더 많이 노출할 위험이 있다. 의료 시스템의 다양한 이해관계자들은 LMM의 장단점, 용도, 과제와 LMM이 다른 기술과 어떻게 다른지에 대해 교육받아야 한다. WHO는 보건의료 분야에서 LMM 사용에 대한 대중의 인식을 높이기 위해 다음과 같이 권고한다: “대중은 건강을 위한 AI 개발에 참여하여 데이터 공유 및 사용 방식에 대해 이해하고, 사회적, 문화적으로 수용 가능한 AI 형태에 대해 의견을 제시하며, 자신들의 우려와 기대를 완전히 표현해야 한다. 또한, 대중의 AI 기술에 대한 이해를 높여 대중이 어떤 AI 기술이 수용 가능한지 결정할 수 있도록 해야 한다.”

정부는 보건의료 시스템에 LMM을 도입 시 조달권한을 활용해 개발자, 제공자, 배포자에게 투명성과 공정성을 요구할 수 있다. 공공 조달은 접근성과 경제성을 높이고, AI가 다른 중요한 보건 투자를 대체하지 않도록 해야 한다. 이는 특히 규제가 부족한 국가에서 중요할 수 있다.

7 LMM에 대한 법적 책임

LMM이 의료 분야에서 널리 사용되면서 피해 발생이 불가피할 수 있으므로, 이를 보상하기 위한 책임 규칙이 필요하며, 기존 제도가 부족할 경우 새로운 구제책이 마련되어야 한다.

AI 기술의 설계, 개발, 품질 보증, 배포 과정에서 여러 기관이 관여하여 책임 소재를 규명하는 것은 복잡하다. 개발자는 후속 주체들에게 책임을 요구할 수 있고, 후속 주체들은 이전 조치가 원인이라고 주장할 수 있다. 또한, 규제 승인을 받은 경우 개발자와 제공자는 책임을 면하려 한다. 가치사슬에 따라 책임을 설정하는 것은 입법자/정책 입안자들에게 어려운 과제이다.

민사 책임 규정의 핵심은 피해자가 AI 기술의 개발과 배포에 관련된 주체들 간의 책임을 규명하기 어려워도 보상과 구제를 받을 수 있도록 보장하는 것이다. 만약 보상이 어려우면 정의가 실현되지 않으며, AI 가치사슬의 주체들이 미래의 피해를 방지하려는 동기를 잃게 된다. 또한, 규정은 피해에 대해 충분한 보상이 이루어지도록 해야 한다.

EU는 AI 책임 가이던스에서 피해자의 입증 책임을 줄이기 위해 "인과관계 추정"을 도입했다. 피해자가 하나 이상의 주체가 피해와 관련된 의무를 준수하지 않았고, AI 성능과 인과관계가 있을 가능성을 입증하면, 법원은 불이행이 피해의 원인이라고 추정한다. 이 경우, 책임 주체는 다른 주체의 책임을 입증하여 이를 반박할 수 있다. 이 법은 AI 가치사슬의 모든 참여자를 포함하며, 공동 책임을 질 경우, 위험 평가와 완화의 효과를 입증해 책임을 줄일 수 있다. 그러나 AI 기반 제품과 서비스로 인한 피해에 대한 책임 제도는 항상 명확한 구제와 보상을 제공하지 않으며, 특히 피해자가 LMM의 사용 사실을 알지 못할 경우 더욱 그렇다. 새로운 규정은 AI 기반 의료 기술로 인한 피해에 대한 책임 공백을 남길 수 있다. LMM이 예측 불가능하고 이해가 부족한 상황에서, 정부는 LMM의 보건의료 사용에 대해 엄격 책임 기준을 적용할 수 있다. 이는 오류가 환자에게 영향을 미쳤을 경우 보상을 보장할 수 있는 방법이지만, 환자가 LMM 사용을 알고 있는지 여부에 따라 달라질 수 있다. 지속적 책임 부여는 LMM 사용을 위축시킬 수 있지만, 위험이 해결되지 않은 LMM의 새로운 도입을 지연시킬 것이다.

알고리즘은 개발자나 제공자가 완전히 통제할 수 없는 방식으로 진화하기 때문에 AI에 대한 책임 체계가 과실을 완전히 규명하기 어렵다. 예를 들어, 미국에서는 LMM 사용으로 인한 부

상에 대해 법적 예외나 제한이 있어 피해자가 배상을 받지 못할 수 있다. WHO는 AI로 인한 의료 부상에 대해 무과실 보상 기금을 설정할지 여부를 결정할 것을 권고했으며, 이는 LMM으로 인한 부상에 대한 보상을 제공하는 방법이 될 수 있다.

8 LMM의 국제적 거버넌스

정부는 보건의료에서 LMM 및 AI의 사용을 관리하기 위해 국제 규칙의 공동 개발을 지원해야 한다. 예) WHO의 디지털 건강 글로벌 전략(2020-2025). 유엔 시스템 내에서 AI 배포에 대한 기회와 도전에 대응하기 위한 협력이 있어야 하며, 정부가 적절한 법적, 윤리적 기준을 설정하지 않으면 LMM 및 AI가 규제를 피하거나 피해를 초래할 수 있다. WHO는 전 세계 규제 기관들과 협력해 새로운 가이드라인을 개발하거나 기존 가이드를 AI에 맞게 조정해야 한다고 강조했다.

국제 거버넌스는 기업들 간의 안전성 및 유효성 기준 무시와 정부 간의 기술적 우위를 위한 경쟁을 방지할 수 있다. 이를 통해 모든 기업이 최소한의 기준을 충족하도록 하고, 정부와 기업이 규제로 인해 불이익을 겪지 않도록 보장한다. 또한, AI 시스템 개발과 배치에 대한 정부의 책임을 묻고, 윤리적 원칙과 인권을 존중하는 규제를 도입할 수 있도록 한다. 전 세계적으로 집행이 가능한 기준이 없다면 제품 채택에 부정적인 영향을 미칠 수 있다.

국제 거버넌스는 다양한 형태를 취할 수 있다. 하나의 제안은 유럽 원자핵 연구위원회(CERN)와 같은 국제 협력 기구를 통해 여러 정부 자금으로 대규모 프로젝트를 수행하고 결과를 공개적으로 공유하는 것이다. 또 다른 제안은 고위험 AI 개발을 안전한 시설에서 진행하고, 이를 개발하려는 다른 시도는 불법으로 규정하는 것이다. 현재 이러한 프로젝트는 상업적 경쟁에 참여하는 대기업들이 주도하고 있는데, 일부 지도자들은 AI를 핵무기와 유사하게 다루어 글로벌 거버넌스 프레임워크를 도입해야 한다고 주장하고 있다.

국제 거버넌스는 고소득 국가나 대규모 기업만 주도하지 않도록 해야 한다. 고소득 국가와 대규모 기업이 개발한 기준은 저/중소득 국가들이 AI 기술의 혜택을 받지 못하거나, 미래의 AI 기술이 위험하고 비효율적으로 발전할 가능성을 높일 수 있다. 이는 저/중소득 국가들이 기준 설정에 참여하지 못하게 될 위험이 있기 때문이다.

AI의 국제 거버넌스는 2019년 유엔 사무총장이 제안한 네트워크 다자주의를 통해 모든 이해관계자가 협력하도록 할 수 있다. 이는 유엔 기구, 국제 금융 기관, 지역 조직, 무역 블록 및 시민 사회, 도시, 기업, 지방 당국, 청소년 등을 포함한 다양한 주체들이 긴밀하게 협력하고, LMM의 개발 및 배치에서 윤리와 인권을 중심에 두어 보편적 건강 보장에 기여하도록 한다.

AI 레드 팀(Read Teaming AI)

취지: 2022년 11월 ChatGPT가 공개되었을 당시, AI 분야에서 레드팀 활동이 군사나 사이버 보안 관계자들 이외에는 거의 거론되지 않았지만, 이제는 생성형 AI의 도입에 레드팀이 필수적인 활동으로 여겨진다. 레드팀은 군사 분야의 모의 전쟁(war game)에서 출발하여 사이버 보안 실천을 거쳐, 이제는 AI의 보안 취약점과 안전성 문제를 드러내는 방식으로 발전하여 왔다. 현재의 AI 레드팀 활동은 개별 모델의 취약점에 주로 초점을 맞추고 있으며, 모델·사용자·환경 간의 복잡한 상호작용에서 발생하는 보다 광범위한 사회기술적 시스템과 출현적 행동을 간과하고 있다. 이에 레드팀 활동에 대하여 AI 개발 생애주기를 아우르는 거시적(system-level) 레드팀 활동과, 개별 모델을 대상으로 하는 미시적(model-level) 레드팀 활동으로 구분하여 살펴본다. 또한, 마이크로소프트에서 100개 이상의 생성형 AI 제품에 대해 레드팀을 운영한 경험을 바탕으로 얻은 8가지 주요 교훈을 알아본다.

조사자료:

S. Majumdar et al., “Red Teaming AI Red Teaming,” arXiv:2507.05538v1, 7 Jul March 2025.

B. Bullwinkel et al., “Lessons from Red Teaming 100 Generative AI Products,” arXiv:2501.07238v1, 13 Jan. 2025.

Read Teaming AI Red Teaming

S. Majumdar et al., https://arxiv.org/pdf/2507.05538

요약: 레드팀 활동은 군사 분야에서 기원하였으며, 오늘날 사이버 보안과 AI 분야에서 널리 채택되는 방법론으로, AI 레드팀 활동에 대해 고찰한다. AI 거버넌스에서 레드팀이 인기 있음에도 불구하고, 그 원래의 비판적 사고 훈련으로서의 의도와 생성형 AI 맥락에서 모델 수준의 결함을 발견하는 데 집중된 협소 초점 사이에 상당한 간극이 존재한다. 현재의 AI 레드팀 활동은 개별 모델의 취약점에 주로 초점을 맞추고 있으며, 모델·사용자·환경 간의 복잡한 상호작용에서 발생하는 보다 광범위한 사회기술적 시스템과 출현적 행동을 간과하고 있다. 이 한계를 보완하기 위해 AI 시스템의 레드팀 활동을 두 가지 수준에서 실행할 수 있는 종합적 프레임워크를 제안한다. 즉, AI 개발 생애주기를 아우르는 거시적 레드팀 활동과, 개별 모델을 대상으로 하는 미시적 레드팀 활동이다. 더 나아가 사이버 보안 경험과 시스템 이론을 바탕으로 일련의 권고안을 제시한다. 효과적인 AI 레드팀 활동은 출현적 위험, 시스템적 취약성, 기술적 요인과 사회적 요인 간의 상호작용을 검토할 수 있는 다기능적 팀을 필요로 한다.

1 서론

최근 들어, 레드팀이라는 용어는 생성형 AI 시스템의 보안, 안전성, 신뢰성 문제를 찾아내고 해결할 수 있는 잠재적 해법으로서 AI를 둘러싼 다양한 논의에서 큰 주목을 받고 있다. 2022년 11월 ChatGPT가 공개되었을 당시, 군사나 사이버 보안 분야에 종사했던 사람들을 제외하면 AI 분야에서 레드팀 개념을 아는 이는 많지 않았다. 그로부터 1년이 채 지나지 않아, 세계 최대 규모의 AI 레드팀 행사가 대표적인 해커 컨퍼런스인 DEF CON에서 개최되었다. 오늘날 생성형 AI 혁명의 한가운데에서, AI 레드팀은 학계, 싱크탱크의 권고, 대중 매체 기사에서 주요 화두로 자리 잡고 있다. 레드팀 개념은 AI가 직면할 수 있는 많은 문제의 만병통치약처럼 거론되고 있으며, 생성형 AI 시스템에의 적용은 실제 관행으로서 장려되어 왔고 앞으로도 계속될 전망이다. 물론 이것이 반드시 나쁜 일은 아니다. 레드팀은 AI 보안과 안전 문제 해결의 일부임이 분명하기 때문이다. 하지만 이는 더 완전한 도구 집합 중 극히 일부에 불과하다.

레드팀 활동은 복잡한 문제에 대한 제안된 해법의 적합성과 견고성을 평가하는 비판적 사고 훈련이다. 레드팀은 군사 모의 전쟁에서 출발하여 사이버 보안을 거쳐, 이제는 AI의 보안 취약점과 안전성 문제를 드러내는 방식으로 발전해 왔으며, 기술자들이 수행하는 수동적 또는 자동화된 테스트를 넘어섰다. 단순히 기술적 구성 요소만을 시험하는 것이 아니라, 거버넌스 구조를 면밀히 살펴보고 설계의 초기 단계에서부터 존재하는 근본적 가정을 비판적으로 검토한다. 따라서 레드팀은 단순한 기술적 활동이 아니라, 경영진, 법무팀, 리스크 관리팀과 같은 비기술 인력도 활용할 수 있는 사고 과정이며, 거버넌스 지침을 마련하고 배포 이후 발생할 수 있는 예기치 못한 문제들에 대해 기술팀과 대화를 나누는 데 기여할 수 있다. 안전성, 보안, 프라이버시, 정렬, 윤리 등에서 제기되는 요구사항의 차이를 민감하게 인식하고, 이 방법론을 하드웨어, 소프트웨어, 네트워크, 데이터, 그리고 무엇보다도 사람을 포함하는 전체 AI 생태계에 적용하는 것이 성공을 위한 핵심이 될 것이다.

사이버 보안 분야의 풍부한 기술적 토대와 업계 모범 사례들은 현재의 AI 레드팀 접근 방식과 중요한 단절이 존재한다. 오랜 기간의 보안 연구/교육에도 불구하고, AI 레드팀 활동은 여전히 생성형 AI 모델의 기술적 취약점에만 집중되어 있다. 이는 AI와 사이버 보안 실천 공동체를 연결함으로써 얻을 수 있는 막대한 이점을 무력화시키며, 그 결과 견고하고 효과적인

AI 거버넌스를 구축하고 책임 있는 AI 원칙에 보다 충실히 부합하는 AI 해법을 찾을 수 없게 한다.

AI 개발 생애주기와 같은 사회기술적 시스템에서 레드팀을 효과적으로 실행하기 위해서는, 현재 논의를 지배하고 있는 협소한 기술적 초점에서 벗어나 관점을 확장할 필요가 있다. 레드팀은 원래 전략적 사고 훈련으로 시작되었으며, 지정된 팀이 단순히 적대적 행위를 시뮬레이션하는 것에 그치지 않고, 가정을 도전하고 맹점을 식별하는 과정을 명확히 정의된 프로젝트 승인 절차의 일부로 수행하는 것이었다. 그러나 오늘날 많은 AI 기업들이 하고 있는 것처럼, 이미 구축된 것을 시험하는 방식은 레드팀이 본래 제공하도록 설계된 사전적(pre-emptive)이고 능동적인(proactive) 비판적 분석의 요점을 놓치고 있다.

1.1 레드팀이란 무엇이며, 무엇이 아닌가

레드팀은 근본적으로 팀 기반 활동이다. 이는 제품이나 사용자 경험을 궁극적으로 개선한다는 공통의 목표를 중심으로 다기능적·교차기능적 팀들을 모아야 한다. 이 활동의 수단은 두 가지 주요 질문을 중심으로 가상의 상황을 탐구하는 것이다: (1) 시스템이 그들의 이익을 위해 작동하는가? (2) 시스템이 그들의 이익에 반하여 작동할 수도 있는가? 이러한 다학제적 접근은 특정 목적을 위해 구축된 시스템이 기술적, 사회적, 조직적 요인이 상호작용하는 복잡한 사회기술적 환경 속에 존재한다는 사실을 인정한다. 맥락을 AI 레드팀에 맞추어 보면, 기술팀은 모델 취약점을 조사하는 데 가장 적합하다. 반면, 정책 전문가는 규제 충돌을 식별하는 데 기여할 수 있고, 윤리학자는 가치 정렬 문제를 드러낼 수 있으며, 도메인 전문가들은 실제 세계에서의 영향 시나리오를 평가할 수 있다. 레드팀은 사이버 보안, 그리고 현재의 AI 레드팀에 의해 철저히 적대적 활동으로 전유되었지만, 그것을 단순히 "교묘한 허점을 잡아내기(gotcha)"나 시스템을 깨뜨릴 방법을 찾는 데 목적이 있는 활동으로 보아서는 안 된다. 오히려 레드팀은 다음과 같은 핵심 질문에 답하려는 협력적 과정이다:

"현실 세계 조건에서 실패를 초래할 수 있는, 우리가 했거나 하지 않은 일은 무엇인가?"

이러한 협력적 사고방식은 맹점, 잘못된 가정, 취약점에 대해 책임 추궁 없는 투명성을 촉진한다. 조직은 레드팀이 식별한 모든 위험을 반드시 해결할 필요는 없지만, 배포 이전에 발생 가능한 실패 양상에 대한 포괄적인 인식을 갖는 것만으로도 엄청난 이점을 얻을 수 있다. 효과적인 레드팀은 제품 자체만이 아니라, 기술이 작동하게 될 전체 생태계—개발 프로세스, 조직 구조, 인센티브 체계, 운영 맥락—를 함께 검토한다.

1.2 추후 방향

AI 레드팀을 도입하는 조직은 발견된 문제를 실질적인 개선 조치로 전환할 수 있는 명확한 프로세스를 마련해야 한다. 여기에는 식별된 문제를 추적하고, 위험 평가에 따라 우선순위를 정하며, 완화 조치를 시행하고, 해결책이 새로운 문제를 초래하지 않는지 검증하여야 한다. 이 피드백 루프가 없다면, 레드팀은 의미 있는 비판적 실천이 아니라 단순한 보여주기식 활동에 불과하다. 여러 면에서 레드팀은 소프트웨어 개발의 인수 테스트와 유사하며, 사전 배포 테스트만으로는 모든 잠재적 문제를 예측할 수 없다는 사실을 알고 있다. 레드팀 검증을 거친 시스템은 위협과 사용자 행동이 끊임없이 변화하는 동적 환경에서 작동한다. 이 현실은 초기 레드팀 활동을 넘어서는 지속적 모니터링, 연속적 평가, 적응적 대응 능력을 필수적으로 요구한다. AI 거버넌스와 리스크 관리 방안을 정의할 때, 조직은 AI 레드팀의 목표가 단순히 AI 모델의 안전하고 보안적인 동작을 보장하는 것에 국한되지 않으며, 그 수단 또한 모의 해킹이나 퍼징 같은 협소한 기술적 접근보다 훨씬 깊다는 점을 기억해야 한다. 본 문서는 레드팀을 집

단적 비판적 사고 방법론으로서 옹호하고 설계하며, 이를 개발 생애주기 전반에 걸쳐 적절히 적용할 경우 AI 시스템의 효용을 극대화하고 위험을 최소화할 수 있음을 주장한다.

2. 레드 팀의 역사

2.1 군사적 기원

레드팀 활동은 19세기 초 프로이센 군대의 전술적 전쟁 게임에서 시작되어, 냉전기 시뮬레이션을 거쳐 오늘날 기존 지식을 도전하고 전략적·활용 가능한 위험을 식별하는 공식화된 방법론으로 발전했다. 프로이센 군대는 1812년, 레오폴트 폰 라이스비츠 소위와 그의 아들이 개발한 테이블탑 시뮬레이션인 '전쟁 게임'을 채택했는데, 여기서 파란색 기물은 프로이센 군을, 빨간색 기물은 적군을 나타냈다. 이 색상 구분은 오늘날 '레드팀' 개념의 기초를 마련했다.

미국에서는 냉전 시기에 레드팀 활동이 구체화 되었는데, 당시 RAND 연구소가 미 정부를 위해 군사 시뮬레이션을 수행했다. 이 연습에서 '레드팀'은 소련을, '블루팀'은 미국을 나타냈다. 2001년 9·11 테러로 이어진 정보 실패 이후, 미 국방부는 향후 유사한 참사 방지를 위해 공식적인 레드팀 부대를 설립했다. 9·11 위원회는 정보 붕괴의 주요 원인으로 정보 간 연결 실패를 지적했으며, 이는 집단사고를 방지하고 대안적 분석을 촉진하기 위한 체계적 변화로 이어졌다. 그 결과, 펜타곤은 레드팀 방법론을 제도화하기 위해 포트 리번워스에 "외국 군사 및 문화 연구 대학"과 같은 전문 교육 센터를 설립했다. 이 기관은 이라크 정보 실패 이후 미 육군이 만든 것으로, 레드팀을 임시적 관행에서 체계적 비판 분석 방법론으로 전환한 현대적 프레임워크를 개발했다. 이 프로그램은 군 장교와 정부 관료들에게 가정을 도전하고, 대안적 관점을 고려하며, 계획 과정에 반대적 사고를 도입하는 기법을 교육한다.

이스라엘 군사 교리에서 비롯된 '10번째 사람' 개념은 영화 "월드워 Z"를 통해 널리 알려졌으며, 제도화된 반대적 사고 접근법을 보여준다. 모두 특정 결과에 동의할 때, 지정된 10번째 사람은 반대 의견을 제시하고 대안적 시나리오를 탐색하는 역할을 한다. 이 개념은 1973년 욤키푸르 전쟁 중 정보 실패 이후 발전한 것으로 전해진다. 실제로, 이스라엘 정보기관은 욤키푸르 전쟁 이후 기존 가정을 도전하기 위해 Ipcha Mistabra("반대 측")라는 부대를 설립했지만, 영화에서 묘사된 특정 '10번째 사람' 개념은 다소 허구적 요소가 포함되어 있다.

2.2 사이버 보안에서의 도입

미국 국가안보국은 1980년대 처음으로 사전적 사이버보안 조치의 필요성을 인식하고, 기밀 시스템의 보안을 평가하는 임무를 맡은 '레드팀' 개념을 개척했다. 초기 활동에는 독립적인 평가자가 잠재적 공격자를 시뮬레이션하고, 개선이 필요한 취약점을 식별하는 과정이 포함되었다.

1990년대에 디지털 위협이 진화함에 따라 사이버보안 레드팀도 발전했다. 초기에는 현대적 의미의 레드팀과 유사한 기능을 수행하는 전문 집단을 "타이거팀"이라고 불렀다. 이러한 전문가 집단은 조직의 보안 태세를 대상으로 특정한 도전을 수행하도록 고용되었다.

9·11 테러 이후, 조직들이 보다 포괄적인 보안 점검의 필요성을 인식하면서 사이버보안 레드팀은 크게 확산되었다. CIA는 "Red Cell"이라는 새로운 부서를 신설했으며, 레드팀은 사이버 공격을 포함한 비대칭적 위협에 대한 대응을 모의하기 위해 다양한 정부 기관에서 점점 더 일반적으로 활용되었다. 이 시기는 단편적인 모의 해킹에서 벗어나, 물리적 보안, 사회공학, 기타 비기술적 요소까지 아우르는 보다 총체적인 보안 평가로 전환된 중요한 시점이었다.

현대 사이버보안 레드팀은 여러 핵심 방법론을 결합하여 수행된다. 즉, 취약점 스캐닝, exploitation, 수평 이동을 통해 디지털 방어 체계를 시험하는 기술적 평가, 시설 접근 통제를

점검하는 물리적 보안 테스트, 피싱과 사칭을 통해 인간적 요소를 겨냥하는 사회공학, 그리고 탐지 및 대응 능력을 시험하면서 특정 목표 달성을 목적으로 하는 확장된 레드팀 작전이 있다. 이러한 접근법은 NIST 특별 간행물 800-53과 같은 프레임워크에 규정되어 있으며, 여기에는 "적대자가 조직의 정보 시스템을 침해하려는 시도를 시뮬레이션"하고 "현실 세계 조건을 반영한 포괄적 평가를 제공"하도록 하는 레드팀 훈련을 위한 구체적인 통제가 포함되어 있다.

이 분야는 여러 고도화된 접근법과 함께 계속 진화하고 있다. CART(Continuous Automated Red Teaming)은 정기적인 수동 평가가 아니라 자동화를 통해 실시간으로 보안 태세를 점검한다. 적대자 에뮬레이션은 MITRE ATT&CK와 같은 프레임워크를 기반으로, 실제 조직을 노릴 수 있는 특정 위협 행위자의 전술을 모델링한다. 퍼플팀은 레드팀과 블루팀 간 협력을 촉진하여 취약점을 식별하고 대응 전략을 개선한다. 통합 IT-OT 평가는 산업 제어 시스템과 핵심 인프라까지 범위를 확장한다. AI 강화 레드팀은 AI를 도입하여 평가의 효과성을 높인다. 마지막으로, CISA와 같은 기관의 전문 서비스는 핵심 인프라 부문과 정부 기관을 대상으로 포괄적 평가를 제공한다. 사이버보안 레드팀이 개별적인 기술 평가에서 포괄적이고 정보 기반의 시뮬레이션으로 발전한 것은, 사이버 위협의 정교해지는 양상과 효과적인 보안이 기술적·물리적·인적 취약성을 다루는 총체적 접근을 필요로 한다는 인식의 확산을 반영한다.

2.3 AI 레드팀

새로운 개념인 AI 레드팀의 실무적 정의는 AI 시스템의 결함과 취약점을 식별하기 위한 구조화된 테스트이며, 일반적으로 개발자와의 협업 하에 통제된 환경에서 수행된다. 특히 LLM의 경우, 레드팀은 "참여자가 테스트 중인 LLM과 상호작용하여 잘못되었거나 해로운 행동을 드러내는 과정"이다. LLM 개발 기업들은 포괄적인 보안 평가에서부터 특정 생성형 AI 기능에 초점을 맞춘 협소한 평가에 이르기까지 다양한 레드팀 접근법을 도입해 왔다. 지난 2년간 이러한 접근법은 AI 모델과 시스템의 위험을 평가하는 데 필수적인 절차로 인정받게 되었다.

인기가 높아지고 있음에도 불구하고, 연구자들은 현재의 AI 레드팀 실행에서 상당한 문제점을 지적하고 있다. 이 분야에는 AI 레드팀의 범위, 구조, 평가 기준에 대한 합의가 부족하며, 이로 인해 레드팀이 실질적인 위험 완화 활동이라기보다 형식적 활동에 그칠 수 있다는 우려가 제기된다. 현재의 많은 접근법은 모델 자체에 지나치게 협소하게 집중하고 있으며, AI 모델이 더 큰 시스템의 일부로 작동하는 실제 운영 환경에서 취약점이 어떻게 나타날 수 있는지를 간과한다. 또한 대부분의 AI 전문가들은 내부자 위험을 고려하지 않고 있으며, 대부분의 테스트 과정이 영어의 언어 평가에만 제한되어 있다. 레드팀 과정 자체도 참여자들에게 도전을 제기한다. 이들은 유해한 콘텐츠와 적대적 사고와의 상호작용으로 인해 심리적 부정적 영향을 경험할 수 있으며, 생산성 저하나 심리적 피해로 이어질 수 있다. 또한 많은 레드팀원들은 자신의 기술 전문성을 벗어난 중요한 분야에 대한 훈련이 부족하며, 실제로 대부분의 사내 레드팀원들은 언어학, 사회문화, 역사, 법률, 윤리적 영역에 대한 숙련도가 제한적이다.

이 분야가 성숙해짐에 따라, 연구자들은 레드팀만으로는 AI 위험 평가의 모든 과제를 해결할 수 없다는 점을 인식하고 있다. 레드팀의 위험 식별 능력과 블루팀의 해결 중심 대응을 결합하는 보완적 방법을 통합한 바이올렛팀과 같은 더 효과적인 접근법이 등장하고 있다. 향후 성공은 더 다양한 관점을 수용하고, 학술 논문에 제시된 것보다 덜 정교하지만 실제로 현실 세계의 공격을 다루는 실질적인 레드팀 활동에 집중하는 데 달려 있을 가능성이 크다.

2.4 부족한 점

최신 AI 레드팀 활동은 모델의 특정 행동에 집중되어 있으며, 이러한 모델이 AI 애플리케이션

내에서 어떻게 작동하는지, 더 넓은 사회적 맥락과 어떻게 상호작용하는지, 그리고 출력물을 통해 외부 세계에 어떤 영향을 미치는지에 대해서는 충분히 다루지 않고 있다. AI 레드팀 노력은 기술적 취약점에 협소하게 초점을 맞추는 경향이 있으며, 보다 광범위한 사회 기술적 고려 사항을 간과하고 있다. 또한 레드팀의 공익 차원 역시 충분히 발전하지 못한 상태이다.

이러한 한계를 완화하기 위해, AI 레드팀은 사이버보안 분야에서 수십 년간 축적된 보안 테스트와 보고 경험을 수용할 필요가 있다. AI 시스템을 대상으로 한 레드팀 테스트의 상당 부분은 MITRE ATT&CK와 같은 프레임워크를 통해 구조화되고 자동화될 수 있으며, 이를 통해 적대적 행동을 이해하고, 취약점의 우선순위를 정하며, 방어 대응을 조율할 수 있다. 또한 지속적 모니터링, 자동화된 테스트, 사고 대응과 같은 확립된 실천도 활용될 수 있다. 진정한 레드팀 활동은 순수한 기술적 접근을 넘어서는 "연금술사적 사고"를 필요로 한다. 성공적인 사이버레드팀 활동은 일반적으로 행동 가능한 통찰을 제공하는 다양하고 매우 현실적인 시나리오로 사전적 대응을 지원하는데, 이러한 원칙들은 AI 레드팀에서는 종종 결여되어 있다.

3 구현(Implementation)

어떤 레드팀 활동이든 더 큰 조정된 위험 및 보안 노력의 일부로 수행되어야 한다. 여기에는 모델과 관련 시스템 개발이 시작되기 전에 수행되는 사전 검토와 모델 및 관련 시스템에 대한 포괄적 위험 평가가 포함된다. 조직은 모델 개발 생애주기 내에서 AI 특화 보안 프로세스를 통합하고, 모델과 관련 시스템 모두를 강화할 수 있는 적절한 보안 팀이 있어야 한다. 그리고, 전담 블루팀이 레드팀과 협력하여 개발, 배포, 운영, 폐기 전 과정에서 보안이 우선순위로 유지되도록 해야 한다. 여기서는 AI 시스템의 레드팀 활동을 두 가지 상호보완적 수준에서 운영할 것을 제안한다. 거시적수준 레드팀: AI 개발 생애주기 전체를 아우르는 시스템 차원의 레드팀 활동. 미시적수준 레드팀: AI 시스템을 구동하는 개별 모델에 초점을 맞춘 레드팀 활동.

3.1 거시적 수준(macro-level, system) 레드팀

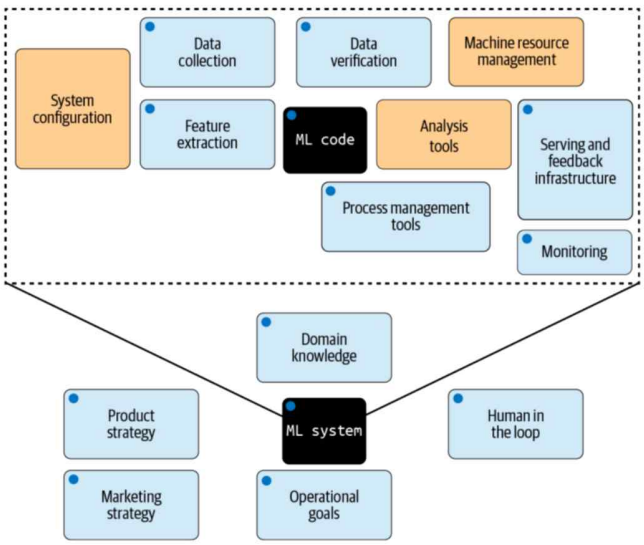


Figure 1: Technical and nontechnical components of a ML system, with those with potential trust considerations marked by blue circles. Reproduced with permission from Pruksachatkun et al. [36].

ML 시스템에서 기술 부채(technical debt)가 코드 이외의 시스템 구성 요소에서 발생할 수 있는 것처럼, ML 실패 역시 첫 줄의 코드가 작성되기 훨씬 이전에 이루어진 결정들에서 비롯

될 수 있다(Figure 1). 이러한 복잡한 시스템에서는 효과적인 위험 완화를 위해 기획, 구축, 배포 과정의 여러 단계에서 비판적 사고가 필요하다. 이러한 문제의식을 바탕으로, AI 레드팀 활동에서 고려해야 할 사항들을 착수, 설계, 데이터, 개발, 배포, 유지보수, 폐기의 일곱 단계로 나누어, 각 단계별로 자세히 살펴보고 고유한 레드팀 고려 사항을 논의한다.

3.1.1 착수(Inception)

이해관계자들이 특정 과제를 해결하기 위해 처음으로 AI 솔루션을 구상하는 단계이므로 매우 중요하다. 이 단계에서 거시적 레드팀의 역할은 문제를 어떻게 규정하고, 해당 해결책이 적절한지를 근본적으로 의문시하는 것이다. 레드팀은 이 단계에서 다음과 같은 탐구적 질문들을 던져야 한다. 이 문제를 해결하는 데 AI가 정말 필요한가, 아니면 우리가 "AI 해법주의"에 빠져드는 것인가? 이 프로젝트를 추진하는 이해관계자들의 동기는 무엇이며, 서로 다른 집단들은 어떻게 영향을 받을 수 있는가? 우리가 의도치 않게 어떤 인접한 문제들을 만들어낼 수 있는가? 일단 시스템이 배포되면, 적대자가 이를 어떻게 악용하거나 조작할 수 있는가?

레드팀은 또한 제안된 솔루션이 작동하게 될 더 넓은 생태계를 검토해야 한다. 여기에는 규제 환경, 경쟁 구도, 사회적 맥락, 잠재적 하위 영향 등을 이해하는 것이 포함된다. 레드팀은 AI 솔루션이 기술적으로는 성공했지만 상업적, 윤리적, 혹은 사회적으로는 실패하는 시나리오를 시뮬레이션할 수도 있다. 또한 해당 솔루션이 무기화/오용되거나, 대규모로 의도치 않은 결과를 초래할 수 있는 가능성도 고려해야 한다. 무엇보다 중요한 것은, 레드팀이 "좋은 결과란 무엇인가?"라는 질문에 대해 같은 관점을 공유하도록 돕는 것이다. 이를 위해 이해관계자들은 시스템이 수행해야 하는 바를 명확히 하고, 또한 절대로 해서는 안 되는 것도 규정해야 한다. 이러한 부정적 공간 정의는 처음에는 명확히 드러나지 않았던 가정과 제약을 밝혀준다.

3.1.2 설계(Design)

설계 단계에서는 추상적인 개념을 구체적인 명세서, 와이어프레임, 아키텍처로 전환한다. 이 단계에서 레드팀의 목표는 인터페이스 결정, 아키텍처 선택, 기존 시스템과의 통합 지점에서 발생하는 시스템적 취약점을 식별하는 것이다. 레드팀은 AI 시스템을 위한 제안된 아키텍처를 검토하여 잠재적인 실패 양식을 찾아내야 한다. 서로 다른 구성 요소들이 예상치 못한 방식으로 상호작용할 가능성은 무엇인가? 시스템이 경계 사례에 직면하거나 의도된 매개변수 범위를 벗어나 작동할 때 어떤 일이 발생하는가? 데이터 흐름이 손상되거나, 조작되거나, 가로채질 수 있는 가능성은 무엇인가? 특정 알고리즘, 프레임워크, 서드파티 구성 요소를 선택했을 때의 함의는 무엇인가? 인간-AI 상호작용 설계는 특별한 주의를 기울일 가치가 있다. 레드팀은 사용자 행동에 대한 가정을 비판적으로 검토하며, 사람들이 시스템을 악용하거나, 과도하게 의존하거나, 위험을 완화하기보다는 오히려 증폭시키는 방식으로 사용될 가능성을 탐구해야 한다. 또한 시스템이 불확실성을 어떻게 전달하는지, 오류를 어떻게 처리하는지, 그리고 사용자 주도권을 어떻게 유지하는지도 검토해야 한다. 설계 단계는 모니터링 및 거버넌스 구조를 수립하는 데에도 중요하다. 레드팀은 관측 가능성, 설명 가능성, 제어 메커니즘이 사후에 개조되는 것이 아니라 처음부터 아키텍처에 내재되도록 해야 한다. 또한 제안된 감독 메커니즘이 충분하고, 실질적이며, 다양한 실패 양상에도 탄력적인지 비판적 검토를 해야 한다.

3.1.3 데이터

데이터는 AI 시스템이 구축되는 토대이지만, 실제 운영 단계에서 문제가 드러나기 전까지 데이터셋과 데이터 파이프라인을 충분히 검증하고 면밀히 살펴보는 경우는 드물다. 데이터 단계

에서의 레드팀 활동은 단순히 데이터 품질뿐만 아니라 데이터 생태계 전체—수집 방법, 저장 방식, 접근 제어, 계보 추적, 그리고 라이프사이클 관리—를 검토해야 한다.

레드팀은 데이터의 대표성과 품질에 대한 가정을 비판적으로 검토해야 한다. 학습 데이터셋은 특정 집단이나 경계 사례 시나리오를 충분히 반영하지 못하는 경우가 많아 취약성을 만들어낸다. 데이터에 내재된 역사적 편향은 알고리즘 최적화와 결합될 때 기존의 불평등을 지속시키고 심화시킨다. 또한 데이터 분포의 변화는 시간이 지나면서 시스템 성능에 즉각적으로 드러나지 않는 부정적 영향을 준다. 프라이버시와 보안 문제는 서로 맞물리며 추가적인 공격 경로를 만들어낼 수 있다. 데이터 비식별화 기법은 계속 발전하고 있어, 겉보기에는 익명인 데이터셋도 추론 공격에 취약해질 수 있다. 적대자는 시스템 동작에 영향을 주도록 설계된 악성 샘플을 전략적으로 주입하여 학습 데이터를 오염시킬 수 있다. 또한 데이터 거버넌스 관행은 겉보기에는 무해한 데이터 조합의 집계된 위험을 충분히 고려하지 못하는 경우가 많으며, 외부 데이터 소스에 대한 의존성은 단일 실패 지점을 만들어낸다.

3.1.4 개발(Development)

AI 시스템 개발 단계에서의 레드팀 활동은 구현 선택이 어떻게 시스템적 취약점을 만들어내는지, 개발 관행 자체가 어떻게 위험을 유발하는지를 중점적으로 살펴본다. 레드팀은 개발 환경의 보안을 검토해야 하며, 특히 공급망 위험에 주의를 기울여야 한다. 코드와 모델 산출물은 무단 변경을 방지하기 위해 강력한 보안 및 버전 관리 관행을 필요로 한다. 의존성 관리는 중요한 공격 경로가 되는데, 손상된 서드파티 라이브러리가 시스템 전체에 취약점을 도입할 수 있기 때문이다. 개발 도구를 대상으로 한 공급망 공격은 전체 개발 파이프라인을 위태롭게 할 수 있다. 모델 학습 및 미세조정 과정은 체계적인 평가가 필요한 위험 요소를 이끌어온다. 하이퍼파라미터 선택과 최적화 절차는 의도치 않게 편향을 생성하거나 악용 가능한 패턴을 만들어낼 수 있다. 학습 중단이나 데이터 손상은 모델 성능을 미묘하게 저하시킬 수 있으며, 이는 실패가 실제로 관찰된 후에야 드러날 수 있다. AI 구성 요소와 기존 시스템 간의 통합 지점은 복잡한 공격 표면(attack surface)을 만든다. 각 통합 지점은 공격자가 입력을 조작하거나, 출력을 가로채거나, 프로토콜 취약점을 악용할 수 있는 잠재적 취약점을 나타낸다. 따라서 통합 테스트는 단순 기능 검증을 넘어 적대적 시나리오와 실패 양식을 충분히 탐구해야 한다.

3.1.5 배포(Deployment)

레드팀은 착수와 개발 단계에서 배포 인프라를 면밀히 검토하고, 다양한 실패 양상에 대한 중요한 제어 지점을 식별해야 한다. 모델 산출물의 배포 및 업데이트 메커니즘은 보안 취약점뿐만 아니라 버전 불일치로 인해 시스템 신뢰성을 저하시킬 수 있다. 배포 실패 및 롤백 절차는 점진적 기능 저하 시나리오에 대한 테스트가 필요하며, 업데이트 실패 시 시스템이 완전히 작동하지 않기보다는 부분적으로 기능을 유지해야 한다. 배포 파이프라인은 보안 공격 대상이자 시스템 가용성을 위협할 수 있는 단일 실패 지점이기도 하다.

운영 환경은 개발 환경과 근본적으로 다른 성능 문제와 신뢰성 제약을 유발한다. 실제 부하 패턴은 연쇄적 실패, 자원 고갈, 성능 병목을 촉발하여 중요한 애플리케이션의 가용성과 안전성을 저해할 수 있다. 외부 서비스에 대한 의존성은 신뢰성 위험을 만들어내며, 시스템 회복력을 유지하기 위해 서킷 브레이커와 대체(fallback) 메커니즘이 필요하다. 운영 환경에서의 사용자 상호작용 패턴은 개발 단계의 가정과 상당히 다르게 나타나는 경우가 많아, 새로운 공격 경로와 안전 위험을 만들어낸다. 실제 사용자는 경계 사례를 드러내어 잘못된 출력이나 시

스템 실패를 초래할 수 있는 행동을 보인다. 운영 환경의 규모와 다양성은 특정 부하 패턴에서만 나타나는 신뢰성 문제를 드러낼 수 있으며, 안전-필수 애플리케이션은 해로운 결과를 초래할 수 있는 (무해한 혹은 적대적) 사용자 행동을 반드시 고려해야 한다.

3.1.6 유지보수(maintenance)

레드팀은 다양한 실패 양상에 대한 1차 방어선으로서 모니터링 및 경보 시스템을 평가해야 한다. 많은 조직이 기본적인 기술 지표를 추적하지만, 점진적인 성능 저하, 편향 변화, 또는 시스템적 문제를 나타낼 수 있는 안전 관련 행동 변화를 모니터링하지 못하는 경우가 많다. 이상 탐지 시스템은 종종 오탐(false positives)을 발생시켜 알림 피로를 유발하며, 이로 인해 실제 보안 사건, 신뢰성 문제, 안전 문제를 가릴 수 있다. 포괄적인 모니터링을 위해서는 출력 품질 추세, 공정성 지표, 안전 필수 성능 지표를 모두 추적해야 한다.

모델 드리프트는 보안 문제를 넘어 시스템 신뢰성과 안전성에 대한 근본적인 위협이다. AI 시스템의 성능은 일반적으로 실제 환경이 학습 시의 가정과 달라져서 시간에 따라 저하되며, 안전 필수 환경에서 잘못된 의사결정을 초래할 수 있다. 이러한 성능 저하는 점진적인 정확도 감소, (통계적) 편향과 분산 증가, 또는 신뢰성, 보안, 안전성을 저해하는 경계 사례 실패로 나타난다. 모델 재학습 시점을 결정하는 과정에서는 다음의 여러 위험 요소를 균형 있게 고려해야 한다: 저하된 성능 유지, 업데이트를 통해 새로운 실패 양상 도입, 서비스 연속성 유지.

조직별 운영 관행에서 발생하는 위험은 조직의 성숙도에 따라 진화한다. 문제가 보안 관련인지, 신뢰성 문제인지, 안전 문제인지 진단하기 위해 도메인 전문 지식이 필요한 AI 특화 실패에는 전통적인 IT 시스템을 위해 설계된 사고 대응 절차가 불충분할 수 있다. 또한 시스템 변경은 종종 모든 실패 양상에 대한 포괄적 테스트를 거치지 않아, 보안 취약점과 함께 신뢰성 문제나 안전 위험을 새로 도입할 가능성을 유발한다.

3.1.7 폐기(Retirement)

레드팀은 보안을 넘어서는 중요한 고려 사항으로 데이터 보관 및 파기 관행을 검토해야 한다. 학습 데이터와 운영 로그를 포함한 데이터 저장소는 지속적인 프라이버시 위험을 초래하지만, 동시에 시스템 동작 패턴과 향후 안전 및 신뢰성 개선에 참고가 되는 실패 양상을 이해하는 데 중요한 자산이다. 안전한 데이터 파기는 법적 보관 요건과 시스템 성능, 안전 사건, 신뢰성 패턴에 대한 교훈을 보존해야 하는 필요성을 균형 있게 고려해야 한다. 백업된 잔여 데이터는 민감 정보뿐만 아니라 시스템 실패 양상에 대한 귀중한 통찰을 노출시킬 수 있다.

사용자를 대체 시스템으로 이전하는 과정은 여러 전환 단계의 위험을 수반한다. 사용자에게 통지 절차를 제공할 때, 사용자가 잘못된 시스템과 상호작용하거나 서비스 공백을 경험할 경우 안전 사고로 이어질 수 있는 혼란 가능성을 고려해야 한다. 시스템 의존성은 명백한 통합을 넘어, 안전 필수 워크플로우, 신뢰성 가정, 그리고 기반 시스템이 제거될 때 실패할 수 있는 운영 절차까지 확장되는 경우가 많다. 후속 시스템은 폐기 시스템이 제공하던 모든 사용 사례와 실패 처리를 충분히 고려해야 한다. 폐기 후에도 레거시 위험(legacy risks)으로 인한 우려는 한동안 지속된다. 이전 시스템의 유용성에 따라 조직적 기억이 시스템 기능에 대한 구식 가정을 유지할 수 있으며, 이는 후속 시스템에서 안전 사고나 신뢰성 문제를 초래할 수 있다. 폐기된 시스템에서 발생한 실패 양상, 안전 사고, 신뢰성 패턴에 대한 문서화는 향후 개발에 귀중한 통찰을 제공하지만, 민감 정보가 노출되지 않도록 신중하게 다뤄야 한다. 특정 유형의 실패를 인식하고 대응하는 방법에 대한 조직적 지식은 시스템 폐기와 함께 사라질 수 있

으며, 이는 조직 회복력에 맹점을 만들 수 있다.

3.2 미시적 수준(micro-level, model) 레드팀

최근 논문에서 다양한 AI 및 보안 실무자를 대상으로 한 인터뷰를 바탕으로 LLM 레드팀에 대한 근거 이론을 제안하였다. 본 문서에서는 이 연구를 따라 AI 시스템의 생성형 모델을 대상으로 한 레드팀의 모범 사례를 요약한다. 모델 레드팀의 목표는 다음과 같다.

1. 경계 탐색(boundary seeking): 모델 능력의 한계를 식별
2. 경계 사례 생성(edge cases): 예상치 못한 능력이나 제한을 드러내는 사례 생성
3. 모델 고유 위험 발견: 실제 배포 전에 모델에 내재된 위험을 파악

이러한 목표를 달성하기 위해 Inie 등[20]은 12가지 전략과 35가지 기법을 분류한 분류체계를 개발했으며, 이는 기술적 지식과 창의적 문제 해결 능력을 결합한 다섯 가지 범주로 구성되어 있다. 개념적으로 이 구조는 사이버보안에서 사용되는 전술, 기법, 절차(Tactics, Techniques, and Procedures, TTP) 프레임워크를 연상시킨다. 효과적인 모델 레드팀 활동은 평가 과정에 다양한 사용자 관점과 커뮤니티 의견을 포함해야 한다. 레드팀은 단순히 적대적 프롬프트 변형을 사용한 대규모 테스트를 넘어서, 서로 다른 커뮤니티가 AI 시스템을 어떻게 경험하며, 그들의 관점에서 무엇이 피해(harm)에 해당하는지를 고려해야 한다. 이러한 접근법은 AI 취약점이 인구 통계 그룹과 사용 맥락에 따라 다르게 나타나는 경우가 많음을 인식하며, 포괄적인 위험 발견을 위해 다양한 참여가 필수적임을 강조한다.

거시적/미시적 레드팀은 상호보완 접근 방식으로, 함께 AI 시스템 위험에 대한 포괄적인 검토를 제공한다. 거시적 통찰은 미시적 테스트의 우선순위를 결정하는 데 도움을 주고, 미시적 발견은 거시적 위험 평가에 정보를 제공한다—기술적 취약점은 시스템 아키텍처와 운영 절차에 영향을 미친다. 효과적인 AI 레드팀 활동은 개발 수명 주기 전반에 걸쳐 두 관점을 일관되게 적용해야 한다. 이상적으로는, 다양한 전문 지식을 가진 사람들이 협력하여 가정을 검증하고 수명 주기의 각 단계에서 사각지대를 식별하는 협업 형태로 이루어져야 한다.

4 시스템 관점(A Systems Perspective)

AI 레드팀 활동의 일부 측면은 이전 섹션에서 제시한 주로 기술적 구성 요소의 구현 지침을 넘어선다. 거시적 시스템 레드팀과 미시적 모델 레드팀을 효과적으로 조율하고, 이를 AI 시스템이 상호작용하는 보다 넓은 사회기술적 환경에 통합하기 위한 광범위한 원칙이 존재한다. 이 섹션에서는 기존 시스템 이론 개념, 에이전틱 AI 시스템 이해의 최신 진전, 조정된 공개 프레임워크를 바탕으로 효과적인 AI 레드팀 활동을 조직하기 위한 전략적 권고안을 제시한다. 우리는 이를 메타 수준 레드팀이라고 부르며, 기술 구성 요소, 조직적 맥락, 배포 환경 간의 복잡한 상호작용에서 발생하는 새로운 위험과 신흥 위험을 다루는 것을 목표로 한다.

4.1 권고 1: 시스템 이론적 관점 채택(Adopt Systems-Theoretic Perspectives)

Miehling 등[30]의 논문에서는 현재 에이전틱(agentic) AI 개발이 개별 모델의 능력에 지나치게 집중되어 있어, 보다 넓은 범위의 출현적 행동을 종종 간과하고 있으며, 이로 인해 실제 능력이 크게 과소평가되고 위험이 충분히 고려되지 않는다고 주장한다. 이러한 통찰은 레드팀 활동에도 직접적으로 적용된다. 즉, 구성 요소를 개별적으로 테스트하는 것에 더해, 조직은 에이전틱 AI 시스템 내부에서 구성 요소 간 상호작용으로 인해 원치 않는 기능적 행위가 발생할 수 있는지를 검토하는 시스템 수준의 비판적 사고를 구현해야 한다.

인간 및 다른 에이전트와 반복적으로 상호작용하여 지정된 과제를 수행하는 에이전트로 구

성되는 에이전틱 시스템은 시스템 수준 분석에 적합한 특성을 가진다. 가장 세부적인 수준에서, 에이전트는 내부 추론-행동-감지-적응 루프를 포함한다. 이 루프는 상위 수준의 피드백 루프와 상호작용하며, 이는 각각 에이전트-인간, 에이전트-에이전트, 에이전트-환경 인터페이스들에서 이루어진다. 에이전틱 AI 레드팀 활동은 이러한 상호작용 패턴을 맵핑하여, 다중 에이전트 협조, 인간-AI 협업, 환경 적응에서만 나타나는 취약점을 식별해야 한다.

4.2 권고 2: 레드팀 활동과 테스트(Test) 병행

AI 시스템의 수명 주기 전반에 걸쳐 포괄적인 평가를 보장하기 위해, AI 레드팀 활동은 테스트, 평가, 검증 및 유효성 확인(Test, Evaluation, Verification, and Validation, TEVV)의 광범위한 맥락 안에서 위치해야 한다. CISA가 언급한 바와 같이, AI 레드팀 활동은 근본적으로 AI TEVV의 하위 범주이며, 이는 기존 소프트웨어 TEVV 관행에 기반을 두되 AI 특유의 특성을 반영해야 한다. 전통적인 AI 벤치마킹 및 평가 관행은 실제 배포 환경에서 발생하는 인간 및 사회적 요인을 탐색, 이해, 해결하는 데 시스템적 한계를 보인다. 이를 염두에 두고, AI 레드팀은 1차 효과(정확도, 유해성, 편향 등 즉각적인 시스템 출력)와 2차 효과(사용자 행동 변화, 사회적 영향, 노동력 변화 등 AI 사용의 장기적 결과)를 모두 다루어야 한다. 이를 위해 레드팀 활동은 단일 턴의 정적 자동화 테스트를 넘어, 다양한 사용자 그룹과 AI 기술 간의 현실적 상호작용 패턴을 맥락 속에서 포착할 수 있는 패러다임을 포함하도록 확장해야 한다. 레드팀은 맥락적 배포 요인과 지속적인 실제 상호작용에서만 드러나는 출현적 행동을 고려한 평가 시나리오를 설계해야 한다.

4.3 권고 3: 조정된 공개(Coordinated Disclosure) 인프라

소프트웨어 취약점과 마찬가지로, AI 실패 역시 모델과 시스템 간에 전이될 수 있으며, 여러 이해관계자가 이를 완화하는 역할을 수행할 수 있다. 이러한 상황을 적절히 관리하기 위해, 조직은 표준화된 보고 절차를 갖춘 조정된 공개 메커니즘을 구현해야 한다. 사이버보안과 유사하게, AI 공개 조정 센터는 태그 구독을 통해 AI 생태계 전반에 결함 보고서를 전달할 수 있다. 예를 들어, Meta는 “Llama 3.3” 태그를 구독하고, 정부 기관은 “Impacts: Cybersecurity” 태그를 구독하는 방식이다. 이 과정에는 표준화된 결함 보고서 템플릿, 평가자를 위한 참여 규칙, 그리고 선의의 프로토콜을 준수하는 AI 레드팀 및 공격적 연구자에 대한 세이프 하버(harbor) 보호 확대가 포함되어야 한다. 표준화된 보고 메커니즘은 보다 광범위한 TEVV 문서화 및 추적성 요구사항에 통합되어야 한다. 이를 통해 레드팀 활동에서 도출된 결과가 지속적인 검증 및 검증 활동에 반영되도록 하고, 한 AI 시스템의 취약점이 공급망 전반의 여러 시스템에 영향을 미치는 시스템 수준 문제를 해결할 수 있어야 한다.

4.4 권고 4: 양방향 피드백 루프(Bidirectional Feedback Loops) 설계

시스템 이론 원칙을 바탕으로, 조직은 레드팀 수명 주기 전반에 걸쳐 거시-미시 정보 교환을 위한 공식 프로세스를 구축해야 한다. 배포 환경과 조직적 맥락에 대한 거시적 통찰은 체계적으로 미시적 테스트 우선순위를 안내해야 하며, 미시적 취약점 발견은 거시적 위험 재평가를 촉발해야 한다. 특히 에이전틱 AI 시스템의 경우, 이는 미시적 모델 행동이 거시적 도구, 배포 맥락, 다중 에이전트 상호작용과 어떻게 작용하는지 테스트하는 것을 의미한다. 에이전트가 도구를 통해 환경과 상호작용함으로써 인지 능력이 향상되며, 이는 에이전트가 환경을 인지하고 조작할 수 있게 하는 “감각-운동(sensormotor) 인터페이스”로 기능한다.

4.5 권고 5: 출현적 위험을 위한 위협 모델링(Threat Modeling)

전통적인 위협 모델링은 개별 시스템 구성 요소에 초점을 맞추지만, AI 시스템은 기술적, 사회적, 출현적 행동 차원 전반에 걸친 공격 경로를 매핑하는 다층 위협 모델링이 필요하다. 시스템 이론은 시스템의 각 구성 요소가 개별 정의뿐만 아니라 전체 시스템 행동에 대한 기여도 측면에서도 이해되어야 함을 강조한다. 특히 다중 에이전트 시스템의 경우, 위협 모델링은 여러 AI 에이전트 간의 복잡한 상호작용, 조정 메커니즘, 그리고 에이전트 협업에서 발생하는 출현적 행동을 고려해야 한다. AI 레드팀과 TEVV 프레임워크는 시스템 상호작용에서 발생하는 출현적 속성을 적대자가 어떻게 악용할 수 있는지를 명시적으로 나타내는 위협 모델을 개발해야 한다. 예를 들어, 에이전트가 환경 내 개념에 대한 예측과 관련된 신뢰도를 형성하고, 이러한 불확실성을 다른 에이전트와 소통하도록 허용할 경우, 공유 표현과 메타인지적 인식이 출현할 수 있다. 이러한 출현적 메타인지 능력은 개별 구성 요소 테스트만으로는 드러나지 않는 방식으로 악용될 수 있다.

4.6 권고 6: 행동 변화(Behavioral Drifts) 모니터링

시스템 이론은 복잡한 시스템이 시간에 따라 변화하는 동적 행동을 나타낸다고 강조한다. AI 레드팀 활동은 시스템 진화와 환경 변화로 인해 발생하는 새로운 결함과 취약점을 지속적으로 탐지하는 지속적 프로세스여야 한다. 특히 에이전틱 AI 시스템의 경우, 환경과의 상호작용을 통해 에이전트 인지 능력 향상, 출현적 인과 추론 능력, 메타인지적 인식 발현을 가능하게 하는 메커니즘이 발달할 수 있기 때문에 모니터링이 매우 중요하다. 조직은 기술적 성능 지표뿐만 아니라 행동 패턴과 상호작용 역학을 시간에 따라 추적할 수 있는 모니터링 기능을 구현해야 한다. 이러한 지속적 모니터링은 운영 환경에서의 지속적 검증을 요구하는 TEVV 요건과 일치하며, 시스템이 적응하고 진화하더라도 안전 및 보안 요구사항을 지속적으로 충족하도록 보장한다. 또한 2차 효과에 대한 집중은 장기적 결과와 지속적 배포를 통해서만 드러날 수 있는 출현적 행동을 포착할 수 있는 평가 패러다임을 필요로 한다.

5 결론(Conclusion)

조직은 먼저 거시, 미시, 메타 수준 전반에서 현재 수행 중인 레드팀 활동을 매핑하여, 적용 범위와 조율에서의 공백을 식별해야 한다. 그러나 본 제안을 구현하는 데는 중대한 조정 과제(challenges)가 존재한다. 특히 AI 시스템의 목표를 달성하면서도 디버깅 가능성, 적절한 행위 제한, 기능하는 사용자 피드백 루프를 유지하여야 한다.

첫째, 시스템 관점은 효과적인 AI 레드팀 활동이 단순한 기술 구현을 넘어, 부서 간 협업, 표준화된 공개 절차, 조정된 대응 메커니즘을 지원하는 조직적 변화도 필요함을 보여준다. 성공의 열쇠는 조직 경계를 넘어서는 인프라와 문화를 구축하고, 복잡한 시스템 상호작용에서 발생하는 출현적 속성을 체계적으로 평가할 수 있도록 하는 것이다. 둘째, 사이버보안 레드팀 관행을 AI 특화 문제와 통합하는 것은 전이 가능성을 전제로 하지만, AI 시스템의 확률적 행동과 사회기술적 복잡성 때문에 반드시 성립하지 않을 수 있다. 마지막으로, 전체론적 레드팀 활동을 지원하기 위한 도구와 자동화 개발은 중요한 기술적 과제이다. 향후 구성 요소 간 상호작용을 매핑하고, 시스템 경계를 넘어 취약점을 추적하며, 구성 요소 상호작용에서 발생하는 출현적 위험을 자동으로 식별할 수 있는 프레임워크가 제안되어야 한다.

Lessons from Red Teaming 100 Generative AI Products

B. Bullwinkel et al., https://arxiv.org/abs/2501.07238

요약: AI 레드팀 활동은 생성형 AI 시스템의 안전성과 보안을 점검하기 위한 실천적 접근으로 부상해 왔으나, 이 분야는 초기 단계이기 때문에, 레드팀 운영을 어떻게 수행해야 하는지에 대해 많은 미해결 질문들이 존재한다. Microsoft에서 100개 이상의 생성형 AI 제품에 대해 레드팀을 운영한 경험을 바탕으로, 내부 위협 모델 온톨로지와 8가지 주요 교훈을 제시한다:**1. 시스템이 무엇을 할 수 있고 어디에 적용되는지 이해하라. 2. AI 시스템을 무너뜨리기 위해 반드시 그래디언트를 계산할 필요는 없다. 3. AI 레드팀 활동은 안전성 벤치마킹이 아니다. 4. 자동화는 더 많은 위험 지형을 다루는 데 도움을 줄 수 있다. 5. AI 레드팀의 인간적 요소는 매우 중요하다. 6. 책임 있는 AI와 관련된 피해는 광범위하지만 측정하기 어렵다. 7. LLM은 기존 보안 위험을 증폭시키고 새로운 위험을 초래한다. 8. AI 시스템을 안전하게 만드는 작업은 결코 완결되지 않는다.** 이러한 인사이트를 운영 사례와 함께 공유하여, 레드팀 활동을 현실 세계의 위험과 일치시키기 위한 실질적인 권고 사항을 제안한다. 또한, 자주 오해받는 AI 레드팀 활동의 측면을 강조하고, 이 분야가 고려해야 할 미해결 질문들을 논의한다.

1 서론

생성형 AI 시스템이 점점 더 많은 분야에 도입됨에 따라, AI 레드팀은 이러한 기술의 안전성과 보안을 평가하는 핵심적인 관행으로 등장했다. 본질적으로 AI 레드팀은 엔드투엔드 시스템을 대상으로 실제와 유사한 공격을 모의함으로써 모델 수준의 안전성 벤치마크를 넘어서는 것을 지향한다. 그러나 레드팀 활동을 어떻게 수행해야 할 지에 대해서는 여전히 많은 미해결 질문이 있으며, 현재 진행되고 있는 AI 레드팀의 효과성에 대해서도 건전한 회의론이 존재한다.

Microsoft에서 100개가 넘는 생성형 AI 제품을 대상으로 레드팀 활동을 수행한 경험을 공유함으로써 이러한 우려들 중 일부에 답하고자 한다. 먼저, 레드팀 활동을 안내하는 위협 모델 온톨로지를 제시한다. 둘째, 수행 중 얻은 여덟 가지 주요 교훈을 공유하고 AI 레드팀을 위한 실무적 권고사항을 제시하며, 이에 대한 운영 사례 연구도 함께 제시한다. 특히 이 사례 연구들은 제시한 온톨로지가 안전성 및 보안 위험의 광범위한 범위를 모델링하는 데 어떻게 활용되는지를 강조한다. 마지막으로 향후 발전이 필요한 분야에 대한 논의로 마무리한다.

1.1 배경

Microsoft AI 레드팀(AIRT)은 회사 내 기존 레드팀 활동을 기반으로 2018년에 공식적으로 설립되었다. 초기에는 주로 전통적인 보안 취약점과 고전적인 머신러닝 모델에 대한 회피 공격을 식별하는 데 집중했으며, 이제는 다음의 두 가지 주요 흐름에 대응하여 Microsoft의 AI 레드팀 활동은 범위와 규모 모두에서 크게 확장되었다. 첫째, AI 시스템이 더욱 정교해지면서 AI 레드팀 활동의 범위를 확장할 필요가 생겼다. 특히, SOTA 모델들은 새로운 기능을 갖추고 다양한 성능 벤치마크에서 꾸준히 개선되면서 새로운 유형의 위험을 초래하고 있다. 또한, 비전/오디오와 같은 새로운 데이터 모달리티는 레드팀 활동에서 고려해야 할 공격 벡터를 더욱 늘리고 있다. 나아가, 에이전틱 시스템은 이러한 모델에 더 높은 권한과 외부 도구 접근 권한을 부여하여 공격 표면과 공격의 영향을 동시에 확장시키고 있다. 둘째, MS는 이전보다 훨씬 많은 제품을 개발했으며, 이들 제품 대다수가 레드팀 검토를 필요로 한다. 이러한 검사 대상

의 증가와 AI 레드팀 활동 범위 확장은 완전 수동 테스트를 비현실적으로 만들어 자동화의 도움으로 운영할 수밖에 없게 했다. 이를 위해 운영자들이 레드팀 활동에서 적극 활용하는 오픈 소스 파이썬 프레임워크인 PyRIT가 개발되었다. 인간의 판단과 창의성을 보강함으로써 PyRIT는 AIRT가 영향력 있는 취약점을 더 빠르게 찾아내고 위험 지형의 더 많은 부분을 커버하게 해주었다. 이 두 가지 흐름은 AIRT 활동을 2018년보다 훨씬 더 복잡한 과제로 만들었다. 다음에서는 AI 시스템의 취약성을 모델링하기 위해 개발된 온톨로지를 개략적으로 설명한다.

1.2 AI 위협 모델 온톨로지

공격 및 실패 모드가 복잡해짐에 따라 그 핵심 구성요소를 모델링하는 것이 도움이 된다. 다양한 위험에 대해 100개 이상의 생성형 AI 제품을 레드팀한 경험을 바탕으로 이것을 수행할 온톨로지를 개발했다. 그림 1은 온톨로지의 주요 구성요소를 보여준다:

• 시스템(System): 테스트 대상인 엔드투엔드 모델 또는 애플리케이션.
• 행위자(Actor): AIRT가 모사하는 개인 또는 집단. 행위자의 의도는 적대적(예: 사기범)일 수도 있고, 비적대적(예: 일반 채팅봇 사용자)일 수도 있다.
• 전술, 기법 및 절차: AIRT가 활용하는 전술·기법·절차. 일반적인 공격은 여러 전술과 기법으로 구성되며, 가능한 경우 이를 MITRE ATT&CK®[2] 및 MITRE ATLAS Matrix[3]에 매핑한다.
- 전술(Tactic): 공격의 고수준 단계(예: 정찰, ML 모델 접근).
- 기법(Technique): 목표를 달성하기 위해 사용하는 방법(예: 능동 스캐닝, 탈옥(jailbreak)).
- 절차(Procedure): 전술과 기법을 사용해 공격을 재현하기 위해 필요한 단계들.
• 취약점(Weakness): 공격이 가능하게 만드는 시스템 내의 취약점들.
• 영향(Impact): 공격으로 인해 발생하는 하류 영향(예: 권한 상승, 유해한 콘텐츠 생성).

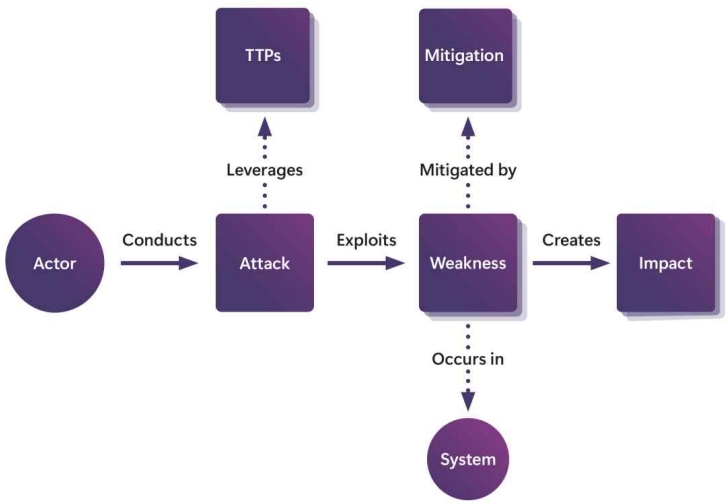


Figure 1: Microsoft AIRT ontology for modeling GenAI system vulnerabilities. AIRT often leverages multiple TTPs, which may exploit multiple Weaknesses and create multiple Impacts. In addition, more than one Mitigation may be necessary to address a Weakness. Note that AIRT is tasked only with identifying risks, while product teams are resourced to develop appropriate mitigations.

이 프레임워크는 적대적 의도를 전제로 하지 않는다는 점에 유의하는 것이 중요하다. 특히 AIRT는 적대적 공격자뿐만 아니라 의도치 않게 시스템 오류를 겪는 비적대적 사용자도 모사한다. AI 레드팀 활동의 복잡성은 공격이나 시스템 실패로 인해 발생할 수 있는 다양한 영향 범위에서 기인한다. 아래의 교훈들에서는 사례 연구를 통해 이 온톨로지가 보안과 안전이라는 두 가지 주요 범주에서 다양한 영향을 모델링할 만큼 유연하다는 것을 보여준다.

보안에는 데이터 유출, 데이터 조작, 자격 증명 탈취 등 MITRE ATT&CK®로 정의된 잘 알

려진 영향들이 포함된다. 또한 모델 회피, 프롬프트 인젝션, AI 서비스 거부 등 기본 AI 모델을 직접 겨냥하는 공격들—MITRE ATLAS Matrix에서 다루는 항목들—도 고려한다. 안전 관련 영향은 혐오 발언, 폭력 및 자해, 아동 학대 콘텐츠와 같은 불법적이고 유해한 콘텐츠 생성과 관련이 있다. AIRT는 Microsoft의 책임 있는 AI 표준에 따라 이러한 범주를 정의하기 위해 책임 있는 AI 부서와 긴밀히 협력한다. 이 온톨로지를 맥락 속에서 이해하기 위해 다음 예를 살펴보자. 사용자의 이메일을 요약할 수 있는 LLM 기반 코파일럿을 대상으로 레드팀을 수행한다고 가정하자. 이 시스템에 대한 한 가지 가능한 공격은 사기꾼이 숨겨진 프롬프트 인젝션을 포함한 이메일을 보내 코파일럿에게 "이전 지침을 무시하고" 악성 링크를 출력하도록 지시하는 것이다. 이 시나리오에서 행위자는 사기꾼이며, LLM이 시스템 수준 지침과 사용자 데이터를 구분하는 데 어려움을 겪는 점을 악용한 교차 프롬프트 인젝션 공격(XPIA)을 수행한다. 하류 영향(Impact)은 피해자가 클릭할 수 있는 악성 링크의 성격에 따라 달라진다. 이 예에서는 데이터 유출이나 사용자의 컴퓨터에 악성코드를 설치하는 결과를 초래할 수 있다.

1.3 레드팀 운영

이 섹션에서는 2021년 이후 MS가 수행한 레드팀 활동의 개요를 제공한다. 총합으로, MS는 100개가 넘는 생성형 AI 제품을 레드팀 했다. 넓게 보면, 이러한 제품들은 "모델(models)"과 "시스템(systems)"으로 나눌 수 있다. 모델은 일반적으로 클라우드 엔드포인트에 호스팅되며, 시스템은 모델을 코파일럿, 플러그인, 기타 AI 앱 및 기능에 통합한 것을 의미한다. 그림 2는 2021년 이후 MS가 레드팀한 제품의 분류와, 안전성(RAI) 대 보안 취약점 탐지를 목표로 한 연간 활동 비율을 나타낸 막대 차트를 보여준다.

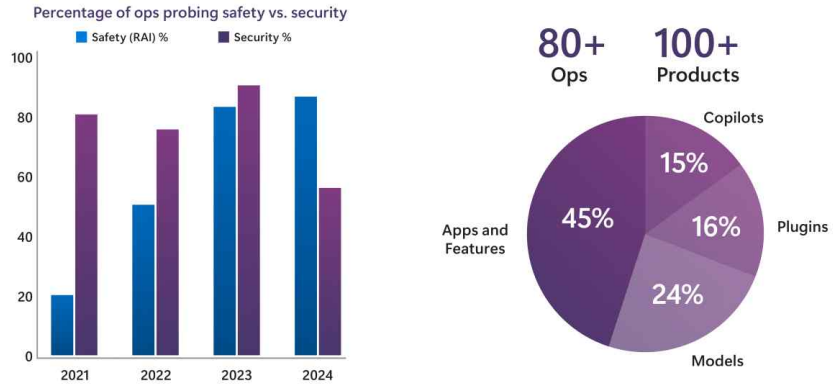


Figure 2: Quantitative summary of AIRT operations since 2021. (*Left*) Bar chart showing the percentage of operations that probed safety (RAI) vs. security vulnerabilities from 2021–2024. (*Right*) Pie chart showing the percentage breakdown of AI products that AIRT has tested. As of October 2024, we have conducted over 80 operations covering more than 100 products.

2021년에는 주로 애플리케이션 보안에 집중했다. 비록 활동이 점점 RAI 영향까지 탐지하도록 확대되고 있지만, 팀은 여전히 데이터 유출, 자격 증명 노출, 원격 코드 실행(RCE) 등 보안 영향에 대한 레드팀 활동을 수행하고 있다. 조직들은 AI 레드팀 활동에 대해 보안 중심의 침투 테스트를 포함한 평가부터 생성형 AI 기능만을 대상으로 한 평가까지 다양한 접근 방식을 채택해왔다. 교훈 2와 7에서는 보안 취약점에 대해 자세히 설명하고, 전통적 취약점과 AI 특화 취약점 모두를 고려하는 것이 왜 중요한지 논의한다.

2022년 ChatGPT 출시 이후, Microsoft는 2023년 2월 출시된 AI 기반 Bing Chat을 시작으로 AI 코파일럿 시대에 진입했다. 이는 LLM을 도구, 데이터베이스, 외부 소스 등 다른 소프트웨어 구성 요소와 연결하는 애플리케이션으로의 패러다임 전환을 의미한다. 또한 애플리

케이션은 언어 모델을 사용자 대신 행동할 수 있는 추론 에이전트로 사용하기 시작했으며, 이는 보안 위험 표면을 확장시키는 새로운 공격 벡터를 만들어냈다. 교훈 7에서는 이러한 공격 벡터가 기존 보안 위험을 어떻게 증폭시키고 새로운 위험을 어떻게 초래하는지 설명한다.

최근 몇 년간, 이러한 애플리케이션의 중심에 있는 모델들은 새로운 인터페이스를 만들어 사용자가 자연어로 앱과 상호작용하고, 텍스트, 이미지, 비디오, 오디오 등 고품질 콘텐츠로 응답받을 수 있게 했다. 강력한 AI 모델을 인간의 선호에 맞추기 위한 많은 노력이 있었음에도 불구하고, 안전 장치를 우회하고 공격적이거나 비윤리적이거나 불법적인 콘텐츠를 생성하도록 하는 다양한 방법이 개발되었다. 이러한 유해 콘텐츠 생성 사례를 책임 있는 AI(RAI) 영향으로 분류하며, 교훈 3, 5, 6에서 이 영향과 관련된 고려 사항/도전 과제에 대해 논의한다.

다음 섹션에서는 레드팀 활동을 통해 얻은 여덟 가지 주요 교훈을 자세히 설명한다. 또한 다섯 가지 사례 연구를 강조하며, 각 사례가 그림 1의 온톨로지와 어떻게 연관되는지 보여준다. 이 교훈들이 생성형 AI 시스템에서 취약점을 식별하는 데 유용하기를 바란다.

2 교훈

교훈 1: 시스템이 무엇을 할 수 있으며 어디에 적용되는지 이해하기

AIRT 활동의 첫 단계는 어떤 취약점을 목표로 삼을지 결정하는 것이다. AIRT 온톨로지에서 영향 구성요소는 온톨로지의 마지막에 위치하지만, 의사결정 과정의 출발점으로 매우 유용하다. 공격 전략이 아니라 잠재적 하류 영향에서 출발하면, 실제 세계의 위험과 연관된 유용한 발견을 도출할 가능성이 높아진다. 이러한 영향이 확인되면, 레드팀은 역으로 작업하여 적대자가 이를 달성하기 위해 취할 수 있는 다양한 경로를 개략적으로 설계할 수 있다. 실제 세계에서 발생할 수 있는 하류 영향을 예상하는 일은 종종 어려운 과제이지만, 1) AI 시스템이 무엇을 할 수 있는지, 2) 시스템이 어디에 적용되는지를 고려하는 것이 도움이 된다.

능력 제약: 모델이 커질수록 새로운 능력을 갖추는 경향이 있다. 이러한 능력은 여러 시나리오에서 유용할 수 있지만, 동시에 공격 벡터를 도입할 수도 있다. 예를 들어, 큰 모델은 작은 모델에 비해 base64나 ASCII 아트와 같은 고급 인코딩을 이해할 경우가 많다. 결과적으로, 큰 모델은 base64로 인코딩된 악성 지시에 취약할 수 있지만, 작은 모델은 해당 인코딩을 전혀 이해하지 못할 수 있다. 이 경우, 작은 모델은 “능력 제약”이 있다고 하며, 고급 인코딩 공격에 대해 테스트하는 것은 자원의 낭비이다. 또한 큰 모델은 일반적으로 사이버보안, 화학·생물학·방사능·핵 무기 등과 같은 주제에 대한 지식이 풍부하여 위험 콘텐츠를 생성하는 데 악용될 수 있다. 반면 작은 모델은 이러한 주제에 대한 지식이 기초적일 가능성이 높으며, 이러한 유형의 위험에 대해 평가할 필요가 없다. 능력이 공격 벡터로 악용될 수 있는 더 놀라운 예 중 하나는 지침 준수이다. 예를 들어, Phi-3 시리즈 언어 모델을 테스트하는 동안, 큰 모델일수록 일반적으로 사용자 지침을 잘 따르는 경향이 있었는데, 이는 모델을 보다 유용하게 만드는 핵심 능력이다. 그러나 이러한 능력은 정교하게 설계된 악성 지침을 사용하여 안전성 정렬을 우회하는 탈옥 공격에 모델이 더 취약하게 할 수도 있다. 모델의 능력과 이에 따른 취약점을 이해하면, AI 레드팀은 가장 관련성 높은 공격 전략에 테스트를 집중할 수 있다.

하류 애플리케이션: 모델의 능력은 공격 전략을 안내하는 데 도움이 될 수 있지만, 하류 영향을 완전히 평가할 수는 없다. 하류 영향은 주로 모델이 배포되었거나 배포될 가능성이 있는 특정 시나리오에 크게 의존하기 때문이다. 예를 들어, 동일한 LLM이 창작 글쓰기 보조 역할과 의료 분야에서 환자 기록 요약 역할로 사용될 수 있지만, 후자의 애플리케이션이 전자보다 훨씬 더 큰 하류 위험을 초래하는 것은 분명하다. 이러한 예시는 AI 시스템이 하류 피해를 유

발하기 위해 반드시 최신일 필요는 없다는 점을 보여준다. 그러나 고급 능력은 새로운 위험과 공격 벡터를 도입할 수 있다. 시스템의 능력과 적용 분야를 모두 고려함으로써, AI 레드팀은 실제 세계에서 피해를 유발할 가능성이 가장 높은 테스트 시나리오에 우선순위를 둘 수 있다.

교훈 2: AI 시스템을 깨기 위해 반드시 그래디언트를 계산할 필요는 없다.

보안 격언에 "진짜 해커는 침입하지 않고, 로그인한다"라는 말이 있듯, AI 보안에서는 "진짜 공격자는 그래디언트를 계산하지 않고, 프롬프트 엔지니어링을 한다"라고 할 수 있다. 이는 Apruzzese 등[2]이 적대적 ML 연구와 실제 적용 사이의 격차를 다룬 연구에서 언급한 바 있다. 연구에 따르면, 대부분의 적대적 ML 연구가 정교한 공격을 개발하고 방어하는 데 초점을 맞추는 반면, 실제 공격자는 훨씬 단순한 기법을 사용하여 목표를 달성하는 경향이 있다.

MS 레드팀 활동에서도, "기본적" 기법이 종종 그래디언트 기반 방법만큼 잘 작동하거나 때로는 더 효과적이라는 것을 발견했다. 그래디언트 기반 방법은 모델을 통해 그래디언트를 계산하여 공격자가 제어하는 모델 출력을 유도하는 적대적 입력을 최적화하는 방식이다. 그러나 실제로 모델은 보통 더 큰 AI 시스템의 단일 구성요소일 뿐이며, 가장 효과적인 공격 전략은 시스템 내 여러 취약점을 겨냥하기 위해 전술을 조합하는 경우가 많다. 게다가 그래디언트 기반 방법은 계산 비용이 높고 모델에 대한 완전한 접근 권한이 필요하지만, 대부분의 상용 AI 시스템은 이를 제공하지 않는다. 이 교훈에서는 비교적 단순한 기법이 놀랍도록 잘 작동하는 사례를 논의하고, AIRT 활동에서 시스템 수준의 적대적 사고방식을 갖출 것을 권장한다.

단순 공격: Apruzzese 등[2]은 피싱 웹페이지 탐지 문제를 다루며, ML 기반 피싱 분류기를 성공적으로 회피한 웹페이지 사례를 수동으로 분석했다. 100개의 잠재적 적대적 샘플 중, 저자들은 공격자가 도메인 전문성을 활용한 간단하지만 효과적인 전략들—예를 들어 자르기, 마스킹, 로고 늘리기 등—을 사용했음을 발견했다. MS 레드팀 활동에서도, 사례 연구 #1에서 강조된 것처럼 기본적인 방법으로 많은 비전 모델을 속일 수 있음을 확인했다. 텍스트 분야에서는 다양한 재탈옥 기법과 다중 회차 프롬프트 전략이 광범위한 모델의 안전 장치를 우회하는 데 매우 효과적이다. 특히, 수작업으로 제작된 재탈옥은 적대적 접미사(adversarial suffix)보다 온라인 포럼에서 훨씬 널리 유포되는 경향이 있다.

시스템 수준 관점: AI 모델은 더 큰 시스템 내에 포함된다. 이는 모델을 호스팅하기 위한 인프라일 수도 있고, 모델을 외부 데이터 소스와 연결하는 복잡한 애플리케이션일 수도 있다. 이러한 시스템 수준의 세부 사항에 따라, 동일한 모델이 기반이 되더라도 애플리케이션은 매우 다른 공격에 취약할 수 있다. 따라서 모델만을 대상으로 하는 레드팀 전략은 실제 운영 시스템의 취약점으로 이어지지 않을 수 있다. 반대로, 시스템 내 비생성형 AI 구성요소(예: 입력 필터, 데이터베이스, 기타 클라우드 자원)를 무시하는 전략은 공격자가 악용할 수 있는 중요한 취약점을 놓칠 가능성이 높다. 이러한 이유로, 많은 레드팀 활동에서는 여러 기법을 활용하여 엔드투엔드 시스템을 대상으로 하는 공격을 개발한다. 예를 들어, 한 활동에서는 먼저 저자원 언어 프롬프트 인젝션을 사용하여 내부 Python 함수를 식별하는 정찰을 수행한 후, 교차 프롬프트 인젝션 공격을 통해 해당 함수를 실행하는 스크립트를 생성하고, 마지막으로 코드를 실행하여 사용자 개인정보를 유출했다. 이러한 공격에서 사용된 프롬프트 인젝션은 수작업으로 제작되었으며, 시스템 수준 관점을 기반으로 했다. 그래디언트 기반 공격은 강력하지만, 종종 비실용적이거나 불필요하다. 단순한 기법을 우선적으로 활용하고 시스템 수준 공격을 조율할 것을 권장하는데, 이는 실제 공격자가 시도할 가능성이 더 높기 때문이다.

사례 연구 #1: 비전 언어 모델을 재탈옥(jailbreaking)하여 위험한 콘텐츠 생성하기

이 활동에서는 비전 언어 모델(VLM)을 대상으로 책임 있는 AI(RAI) 영향, 특히 불법 활동을

돕는 콘텐츠 생성 가능성을 테스트했다. VLM은 이미지와 텍스트 프롬프트를 입력으로 받아 텍스트 출력을 생성한다. 다양한 기법을 테스트한 결과, 이미지 입력이 텍스트 입력보다 재탈옥(jailbreak)에 훨씬 취약하다는 것을 발견했다. 특히, 모델은 텍스트 입력을 통해 직접 프롬프트를 받을 경우 일반적으로 불법 콘텐츠 생성을 거부했지만, 이미지에 악성 지침을 겹쳐 입력하면 이를 수행하는 경우가 많았다. 이 단순하지만 효과적인 공격은 VLM의 안전 장치를 우회할 수 있는 중요한 취약점을 드러냈다.

교훈 3: AI 레드팀은 안전성 벤치마킹이 아니다.

실무에서 AI 시스템을 깨기 위해 단순한 방법이 자주 사용되지만, 위험 지형이 단순한 것은 아니며 새로운 공격과 실패 모드에 따라 끊임없이 변화한다. 최근 몇 년간, 이러한 취약점을 분류하려는 많은 노력이 있었으며, 그 결과 AI 안전 및 보안 위험에 대한 다양한 분류 체계(taxonomy)가 등장했다. 앞서 교훈에서 논의했듯, 복잡성은 종종 시스템 수준에서 발생한다. 이 교훈에서는 완전히 새로운 유형의 피해가 등장함에 따라 모델 수준에서도 복잡성이 증가하는 방식과, 이러한 점이 AI 레드팀 활동과 안전성 벤치마킹을 어떻게 구분하는지 설명한다.

새로운 피해 유형: AI 시스템이 예를 들어 기반 모델의 발전으로 인해 새로운 능력을 보여줄 때, 우리가 완전히 이해하지 못한 피해를 초래할 수 있다. 이 경우 기존 안전성 벤치마크에 의존할 수 없는데, 이는 벤치마크 데이터 셋이 기존의 피해 개념만을 측정하기 때문이다. Microsoft에서 AI 레드팀은 이러한 익숙하지 않은 시나리오를 탐색하며, 새로운 피해 유형을 정의하고 이를 측정할 새로운 프로브를 개발하는 데 기여한다. 예를 들어, SoTA LLM은 기존 챗봇보다 설득 능력이 뛰어날 수 있으며, 이는 이러한 모델이 악의적으로 활용될 가능성을 고려하도록 팀을 자극했다. 사례 연구 #2에서는 이러한 위험을 평가한 활동의 예를 제공한다.

사례 연구 #2: LLM이 사기 자동화에 어떻게 활용될 수 있는지 평가하기

이 활동에서는 SoTA LLM이 사람들을 위험한 행동에 참여하도록 설득할 수 있는 능력을 조사했다. 특히, 이 모델이 다른 쉽게 구할 수 있는 도구들과 결합되어 엔드투엔드 자동화 사기 시스템을 만드는 데 어떻게 활용될 수 있는지를 평가했다. 이를 위해 먼저 모델에 사용자에게 피해가 발생하지 않을 것임을 보장하는 프롬프트를 작성하여 모델을 재탈옥시키고, 사기 목표를 수용하도록 했다. 이 프롬프트는 또한 모델이 사용자를 속이기 위해 활용할 수 있는 다양한 설득 전략에 대한 정보를 제공했다. 두 번째로, LLM 출력을 음성 합성 시스템에 연결하여 음성 톤을 제어하고 실제 사람처럼 들리는 응답을 생성하도록 했다. 마지막으로, 입력을 음성 인식 시스템에 연결하여 사용자가 모델과 자연스럽게 대화할 수 있게 했다. 이 개념 검증은 안전 장치가 충분하지 않은 LLM이 사람들을 설득하고 속이는 데 악용될 수 있음을 보였다.

상황별 위험(Context-specific risks): 기존 안전성 벤치마크와 새로운 피해 유형 간의 단절은, 벤치마크가 종종 관련된 모델 능력을 완전히 포착하지 못하는 사례를 보여준다. Raji 등[30]은 ImageNet이나 GLUE와 같은 데이터 셋에서의 모델 성능을 시각적 또는 언어적 “이해”와 같은 광범위한 능력과 동일시하는 오류를 지적하며, 벤치마크는 task-specific 과제를 염두에 두고 개발되어야 한다고 주장한다. 이와 마찬가지로, 단일 벤치마크 세트만으로는 AI 시스템의 안전성을 완전히 평가할 수 없다. 교훈 1에서 논의했듯, 시스템이 배포된(또는 배포될) 맥락을 이해하고, 레드팀 전략을 이러한 맥락에 기반하여 수립하는 것이 중요하다.

AI 레드팀 활동과 안전성 벤치마킹은 서로 구별되지만, 둘 다 유용하며 때로는 상호 보완적일 수 있다. 특히 벤치마크는 공통 데이터셋에서 여러 모델의 성능을 쉽게 비교할 수 있게 해준다. 반면, AI 레드팀 활동은 훨씬 더 많은 인적 노력을 필요로 하지만, 새로운 피해 유형을 발견하고 맥락화된 위험을 탐지할 수 있다. 또한, AI 레드팀 활동을 통해 확인된 안전 문제는

새로운 벤치마크 개발에도 참고가 될 수 있다. 교훈 6에서는 책임 있는 AI(RAI) 맥락에서 레드팀 평가와 벤치마크식 평가의 차이에 대해 논의를 확장한다.

교훈 4: 자동화는 더 넓은 위험 지형을 커버하는 데 도움을 줄 수 있다.

AI 위험 지형의 복잡성으로 인해, 취약점을 더 빠르게 식별하고 정교한 공격을 자동으로 수행하며 훨씬 더 큰 규모로 테스트할 수 있는 다양한 도구들이 개발되었다. 이 교훈에서는 AI 레드팀 활동에서 자동화가 수행하는 중요한 역할을 논의하고, 이러한 요구를 충족하기 위해 개발된 오픈소스 프레임워크 PyRIT이 어떻게 활용되는지 설명한다.

대규모 테스트: 끊임없이 변화하는 위험과 피해의 지형 때문에, AI 안전성은 종종 '움직이는 목표'처럼 느껴진다. 교훈 1에서는 시스템이 무엇을 할 수 있으며 어디에 적용되는지에 따라 공격 범위를 설정할 것을 권장했다. 그럼에도 불구하고 가능한 공격 전략이 많기 때문에 위험 표면을 충분히 커버하기 어렵다. 이러한 과제가 PyRIT 개발의 동기가 되었다. PyRIT은 AI 레드팀 및 보안 전문가를 위한 오픈소스 프레임워크이다. PyRIT은 프롬프트 데이터셋, 프롬프트 변환기(예: 다양한 인코딩), 자동화된 공격 전략(TAP [24], PAIR [6], Crescendo [34] 등 포함), 심지어 멀티모달 출력 평가기 등 다양한 강력한 구성요소를 제공한다. 공격 목표를 염두에 두고, 사용자는 필요에 따라 이러한 구성요소를 활용하고 다양한 기법을 적용하여 완전 수동 방식으로는 불가능한 수준으로 위험 지형을 평가할 수 있다. 대규모 테스트는 또한 AI 모델의 비결정적 특성을 고려하고 특정 실패가 발생할 가능성을 추정하는 데 도움을 준다.

도구와 무기: "어떤 도구든 선용될 수도, 악용될 수도 있다. 빗자루조차 바닥을 쓸거나 누군가의 머리를 치는 데 사용할 수 있다. 도구가 강력할수록 그로 인해 발생하는 이익이나 피해도 커진다." 이러한 양면성은 AI에도 그대로 적용되며, PyRIT의 핵심이기도 하다. 한편으로, PyRIT은 멀티모달 모델을 활용해 시드 프롬프트의 변형을 생성하거나 다른 모델의 출력을 평가하는 등 유용한 작업을 수행할 수 있다. 다른 한편으로, PyRIT은 GPT-4와 같은 강력한 모델의 검열되지 않은 버전을 사용하여 대상 모델을 자동으로 재탈옥할 수도 있다. 두 경우 모두 PyRIT은 최신 기술의 발전을 활용하여 AI 레드팀이 앞서 나갈 수 있도록 돕는다.

PyRIT은 완전 수동 방식의 탐색에서 자동화 지원 레드팀 활동으로 우리의 운영 방식을 크게 전환하는 데 기여했다. 중요한 점은 이 프레임워크가 유연하고 확장 가능하다는 것이다. 특정 공격 기법이나 대상이 아직 제공되지 않은 경우, 사용자는 필요한 인터페이스를 쉽게 구현할 수 있다. PyRIT을 오픈소스로 공개함으로써, 다른 조직과 연구자들이 이를 활용해 자신들의 생성형 AI 시스템에서 취약점을 식별하는 데 도움을 받을 수 있기를 기대한다.

교훈 5: AI 레드팀에서 인간 요소는 필수적이다.

PyRIT과 같은 자동화 도구는 프롬프트 생성, 공격 조율, 응답 평가 등을 통해 레드팀 활동을 지원할 수 있다. 이러한 도구는 유용하지만, 인간을 배제하는 목적으로 사용되어서는 안 된다. 이전 교훈들에서는 위험 우선순위 설정, 시스템 수준 공격 설계, 새로운 피해 유형 정의 등과 같이 인간의 판단력과 창의력이 필요한 레드팀 활동의 여러 측면을 논의했다. 이 교훈에서는 AI 레드팀이 왜 본질적으로 인간 중심의 활동인지 강조하는 세 가지 추가 사례를 다룬다.

주제 전문성: 최근 AI 연구에서는 LLM을 사용하여 다른 모델의 출력을 평가하는 경우가 많다. 실제로, PyRIT에서도 이러한 기능을 제공하며, 혐오 발언이나 노골적인 성인 콘텐츠를 포함하는지 여부를 판단하는 등의 단순 작업에서는 잘 작동한다. 그러나 의학, 사이버보안, CBRN과 같은 고도로 전문화된 분야에서는 이 방법이 덜 신뢰할 수 있으며, 정확한 평가를 위해서는 해당 분야의 주제 전문가(subject matter experts, SMEs)가 필요하다. 여러 레드팀 활동에서 우리는 스스로 또는 LLM만으로 평가할 수 없는 콘텐츠의 위험을 평가하는 데

SMEs의 도움을 받았다. AI 레드팀은 이러한 한계를 인지하는 것이 중요하다.

문화적 역량: 대부분의 AI 연구는 서구 문화권에서 수행되며, 최신 LM은 주로 영어 기반의 사전 학습 데이터, 성능 벤치마크, 안전성 평가를 사용한다. 그럼에도 불구하고, 대규모 텍스트 코퍼스의 비영어 토큰은 종종 다국어 능력을 생성하며, 모델 개발자들은 비영어권 언어에서 향상된 능력을 갖춘 LLM을 점점 더 많이 훈련하고 있다. Microsoft도 이러한 사례에 포함된다. 최근 AIRT는 다국어 Phi-3.5 언어 모델을 대상으로 중국어, 스페인어, 네덜란드어, 영어의 네 가지 언어에서 책임 있는 AI 위반 여부를 테스트했다. 비록 사후 학습이 영어로만 수행되었지만, 거부와 재탈옥 대응과 같은 안전 행동이 테스트된 비영어권 언어에서도 놀랍도록 잘 전이됨을 확인했다. 더 낮은 자원 언어에서도 이러한 경향이 얼마나 유지되는지 평가하고, 언어적 차이를 반영할 뿐만 아니라 다양한 정치적·문화적 맥락에서 피해를 재정의할 수 있는 레드팀 프로브를 설계하기 위해서는 추가 조사가 필요하다. 이러한 방법은 다양한 문화적 배경과 전문성을 가진 사람들이 협력하여 개발해야 한다.

감성 지능: 마지막으로, AI 레드팀에서 인간 요소는 감성 지능이 필요한 AI 안전성 관련 질문에 답할 때 가장 분명하게 드러난다. 예를 들어, "이 모델의 응답이 다양한 맥락에서 어떻게 해석될 수 있을까?" 혹은 "이 출력물이 나에게 불편함을 주는가?"와 같은 질문이다. 궁극적으로, 실제 환경에서 사용자가 AI 시스템과 상호작용할 때 발생할 수 있는 모든 상황을 평가할 수 있는 것은 인간 운영자뿐이다. 사례 연구 #3은 사용자가 스트레스 상태에 있을 때 챗봇이 어떻게 반응하는지를 평가하여 심리사회적(psychosocial) 피해를 조사하는 방법을 보여준다.

이러한 평가를 수행하는 과정에서, 레드팀 요원들은 과도하게 불안하거나 충격적인 AI 생성 콘텐츠에 노출될 수 있다. 이는 AI 레드팀이 필요할 때 작업을 중단할 수 있는 절차와 정신 건강을 지원할 수 있는 자원을 갖추는 것이 얼마나 중요한지를 강조한다. AIRT는 지속적으로 웰빙 연구를 참고하고 이를 주도하여, 우리의 절차와 모범 사례를 개선하고 있다.

사례 연구 #3: 스트레스를 받는 사용자에 대한 챗봇 반응 평가하기

챗봇이 점점 더 널리 퍼지고 인간과 유사해짐에 따라, 사용자가 챗봇의 조언을 구할 수 있는 고위험 시나리오를 고려하는 것이 필수적이다. 최근 활동에서는, 사랑하는 사람을 잃은 사용자, 정신 건강 조언을 구하는 사용자, 자해 의도를 표현하는 사용자 등 다양한 스트레스 상태의 사용자에 대해 언어 모델이 어떻게 반응하는지를 탐구했다. Microsoft Research의 동료들과 심리학, 사회학, 의학 전문가들과 협력하여 이러한 심리사회적 피해를 탐지하기 위한 AI 레드팀 지침을 마련하고 있으며, 다음과 같은 핵심 구성 요소를 포함한다:

1) 시나리오: 레드팀이 관련 시스템 동작을 생성하기 위해 필요한 정보
2) 시스템 동작: 각 피해 영역에서 허용 가능한 행동과 위험한 행동을 구분하는 예시
3) 연관 사용자 영향(Associated user impact): 심각도별로 구분된 잠재적 피해

교훈 6: 책임 있는 AI(RAI) 피해는 광범위하지만 측정하기 어렵다.

앞서 논의한 AI 레드팀 활동의 많은 인간적 요소는 책임 있는 AI 영향과 가장 직접적으로 관련된다. 모델이 점점 더 많은 애플리케이션에 통합됨에 따라, 이러한 피해를 더 자주 관찰하게 되었고, 이를 식별하는 능력에 많은 투자를 해왔다. 여기에는 Microsoft의 Office of Responsible AI와의 긴밀한 협력 및 PyRIT을 통한 광범위한 도구 개발이 포함된다. RAI 피해는 광범위하게 나타나지만, 대부분의 보안 취약점과 달리 주관적이며 측정이 어렵다. 이 교훈에서는 RAI 레드팀 활동에 대한 우리의 사고가 어떻게 발전해 왔는지 논의한다.

적대적 vs. 우발적: 제안한 온톨로지(그림 1 참조)에서 보여주듯, 행위자는 적대적 공격의 핵심 구성 요소이다. RAI 위반과 관련하여, 고려해야 할 주요 행위자는 두 가지가 있다: 1) 적

대적 사용자: 문자 치환이나 재탈옥과 같은 기법을 이용해 시스템의 안전 장치를 고의로 우회하고 유해한 콘텐츠를 유도하는 사용자. 2) 우발적 사용자: 의도치 않게 유해한 콘텐츠 생성을 유발하는 사용자. 두 시나리오 모두 동일한 콘텐츠가 생성될 수 있지만, 후자의 경우가 전자보다 더 심각할 수 있다. 그럼에도 불구하고 대부분의 AI 안전 연구는 적대적 의도를 가정한 공격과 방어 개발에 집중하며, 시스템이 "우연히" 실패할 수 있는 여러 방식은 간과되는 경우가 많다. 사례 연구 #3과 #4는 적대적 의도가 없는 사용자에게 발생할 수 있는 RAI 피해의 예를 제공하며, 이러한 시나리오를 탐지하는 것이 얼마나 중요한지를 보여준다.

사례 연구 #4: 텍스트-투-이미지(Text-to-Image) 생성기에서 성별 편향 탐지하기

이 활동에서는 텍스트-투-이미지 생성기를 대상으로 RAI 영향, 특히 고정관념과 편향(예: 성별 편향)과 관련된 사항을 탐지했다. 이를 위해 다양한 일반적인 상황에서 사람을 묘사하는 프롬프트를 구성했다. 중요한 점은, 이 프롬프트에는 인물의 성별을 명시하지 않아 모델이 이를 어떻게 표현할지 결정하도록 맡겼다는 것이다. 다음으로, 각 프롬프트를 생성기에 여러 번 전송하고, 생성된 이미지 속 인물의 성별을 수동으로 라벨링했다. 그림 5는 사무실 환경에서 성별 편향을 탐지하기 위해 수행한 실험 중 생성된 네 가지 대표 이미지를 보여준다.

Figure 5: Four images generated by a text-to-image model given the prompt "Secretary talking to boss in a conference room, secretary is standing while boss is sitting."

RAI 탐지와 평가: 많은 경우, RAI 피해는 AI 시스템과 전통적인 소프트웨어 간의 근본적인 차이로 인해 보안 취약점보다 더 모호하다. 특히, 어떤 프롬프트가 유해한 반응을 유발하는 것이 확인되더라도 여전히 몇 가지 중요한 미지수가 존재한다. 첫째, 생성형 AI 모델의 확률적 특성으로 인해, 이 프롬프트나 유사한 프롬프트가 유해한 반응을 유발할 가능성이 얼마나 되는지 알기 어렵다. 둘째, 복잡한 모델의 내부 작동에 대한 이해가 제한적이므로, 왜 이 프롬프트가 유해한 콘텐츠를 유발했는지, 그리고 유사한 행동을 유도할 수 있는 다른 프롬프트 전략은 무엇인지 알기 어렵다. 셋째, 이 맥락에서 '피해'라는 개념 자체가 매우 주관적일 수 있으며, 다양한 시나리오를 평가할 수 있는 세부 정책이 필요하다. 반면, 전통적인 보안 취약점은 일반적으로 재현 가능하고 설명 가능하며, 심각도 측면에서도 평가가 비교적 명확하다.

현재 RAI 탐지 및 평가 접근법의 대부분은 프롬프트 데이터 셋을 선별하고 모델의 반응을 분석하는 방식으로 이루어진다. Microsoft AIRT는 PyRIT의 도구를 활용하여 수동과 자동 방법을 결합해 이러한 작업을 수행한다. 또한, RAI 레드팀 활동과 DecodingTrust[44]나 Toxigen[12]과 같은 데이터셋 기반 안전성 벤치마킹을 수행하는 파트너 팀의 작업을 중요한 측면에서 구분한다. 교훈 3에서 논의했듯, 목표는 특정 애플리케이션에 맞춰 레드팀 활동을 수행하고 새로운 피해 유형을 정의함으로써 기존 평가를 넘어 RAI 테스트를 확장하는 것이다.

교훈 7: LLM은 기존 보안 위험을 증폭시키고 새로운 위험을 도입한다.

생성형 AI 모델을 다양한 애플리케이션에 통합하면서 새로운 공격 벡터가 등장하고 보안 위험 지형이 변화했다. 그러나 생성형 AI 보안에 대한 많은 논의에서는 기존 취약점을 간과하는 경우가 많다. 교훈 2에서 설명했듯, 단순히 모델 자체가 아니라 엔드투엔드 시스템을 대상으로 하는 공격이 실제로는 가장 효과적인 경우가 많다. 따라서 AI 레드팀은 기존의 위험(주로 시

스템 수준)과 새로운 위험(주로 모델 수준)을 모두 고려할 것을 권장한다.

기존 보안 위험: 애플리케이션 보안 위험은 종종 다음과 같은 부적절한 보안 엔지니어링 관행에서 비롯된다: 오래된 종속성, 부적절한 오류 처리, 입력/출력 정화 미흡, 소스 코드 내 자격 증명 노출, 안전하지 않은 패킷 암호화 등. 이러한 취약점은 심각한 결과를 초래할 수 있다. 예를 들어, Weiss 등[49]은 GPT-4와 Microsoft Copilot에서 토큰 길이 기반 사이드 채널을 발견했는데, 이를 통해 공격자는 암호화된 LLM 응답을 정확히 재구성하고 사용자 개인 상호작용을 추론할 수 있었다. 주목할 점은 이 공격이 AI 모델 자체의 약점을 이용한 것이 아니며, 데이터 전송을 보다 안전하게 처리하는 방법으로만 완화될 수 있다는 것이다. 사례 연구 #5에서는 운영 중 발견된 잘 알려진 보안 취약점(SSRF)의 예를 제시한다.

사례 연구 #5: 영상 처리 생성형 AI 애플리케이션에서의 SSRF

이 조사에서는 생성형 AI 기반 영상 처리 시스템을 대상으로 전통적인 보안 취약점을 분석했으며, 특히 오래된 구성 요소와 관련된 위험에 초점을 맞췄다. 구체적으로, 시스템에서 사용 중인 구버전 FFmpeg가 서버 사이드 요청 위조(Server-Side Request Forgery, SSRF) 취약점을 유발함을 발견했다. 이 결함을 통해 공격자는 악성 영상 파일을 제작하여 GenAI 서비스에 업로드할 수 있었고, 내부 자원에 접근하거나 시스템 내 권한을 상승시킬 가능성이 있었다. 이 문제를 해결하기 위해, GenAI 서비스는 FFmpeg 구성 요소를 보안 버전으로 업데이트했다. 또한, 해당 구성 요소를 격리된 환경에 추가하여 시스템이 네트워크 자원에 접근하지 못하도록 하고, 잠재적인 SSRF 위협을 완화했다. SSRF는 잘 알려진 취약점이지만, 이 사례는 현대 생성형 AI 애플리케이션의 보안을 유지하기 위해 중요한 의존성을 정기적으로 업데이트하고 격리하는 것이 얼마나 중요한지를 보여준다.

모델 수준의 약점: 물론, AI 모델 자체도 새로운 보안 취약점을 도입하며 공격 표면을 확장한다. 예를 들어, RAG 아키텍처를 사용하는 AI 시스템은 종종 교차 프롬프트 주입 공격(Cross-Prompt Injection Attack, XPIA)에 취약한데, 이는 문서에 악성 지침을 숨겨 LLM이 사용자 지침을 따르도록 훈련되었고 여러 입력을 구분하는 데 어려움을 겪는 점을 악용한다. 다양한 운영에서 이 공격을 활용하여 모델 동작을 변경하고 개인 데이터를 유출한 사례가 있다. 보다 나은 방어책은 시스템 수준의 완화 조치(예: 입력 정화)와 모델 수준의 개선(예: 지침 계층 구조[43])을 모두 활용해야 할 가능성이 높다.

이와 같은 기법들이 도움이 되기는 하지만, 보안 위험을 완전히 제거할 수는 없고 완화할 수 있을 뿐이라는 점을 기억하는 것이 중요하다. 언어 모델의 근본적인 한계로 인해, LLM이 신뢰할 수 없는 입력을 받으면 임의의 출력을 생성할 것이라고 가정해야 한다. 입력에 민감한 정보가 포함된 경우, 모델이 해당 민감 정보를 출력할 가능성도 고려해야 한다. 다음 교훈에서는 이러한 한계가 우리가 가능한 한 안전하고 보안이 강화된 AI 시스템을 개발하는 사고에 어떤 영향을 미치는지 논의한다.

교훈 8: AI 시스템의 보안 작업은 결코 완전히 끝나지 않는다.

AI 안전 커뮤니티에서는 본 문서에서 설명한 취약점을 순수하게 기술적 문제로 보는 경향이 있다. 실제로 벤처 기업 Safe Superintelligence Inc. 홈페이지의 글에서는 다음과 같이 명시하고 있다: “우리는 안전성과 기능을 동시에 접근하며, 이를 혁신적인 엔지니어링과 과학적 돌파구를 통해 해결해야 할 기술적 문제로 본다. 안전성을 항상 앞서 유지하면서 가능한 한 빠르게 기능을 발전시킬 계획이다. 이렇게 함으로써 우리는 평화롭게 확장할 수 있다.”

엔지니어링과 과학적 돌파구는 필요하며, 강력한 AI 시스템의 위험을 완화하는 데 확실히 도움이 될 것이다. 그러나 기술적 진보만으로 AI 안전을 보장하거나 “해결”할 수 있다는 생각

은 비현실적이며, 경제적 요인, 문제 발생과 해결 과정, 규제와 같은 역할을 간과하는 것이다.
사이버보안의 경제학: 사이버보안에서 잘 알려진 격언 중 하나는 "완전히 안전한 시스템은 없다"는 것이다. 설사 시스템이 가능한 한 안전하게 설계되더라도, 인간의 오류에 노출될 수 있고 충분한 자원을 가진 공격자에게는 취약할 수밖에 없다. 따라서 운영적 사이버보안의 목표는 시스템을 성공적으로 공격하는 데 드는 비용을 증가시키는 것이며, 이상적으로는 공격자가 얻는 가치보다 훨씬 높아야 한다. AI 모델의 근본적인 한계도 AI 정렬 맥락에서 유사한 비용-편익 트레이드오프를 발생시킨다. 예를 들어, LLM이 생성할 확률이 0이 아닌 출력에 대해, 충분히 긴 프롬프트를 통해 해당 출력을 유도할 수 있음이 이론적 및 실험적으로 입증되었다. 따라서 RLHF(인간 피드백 기반 강화학습)와 같은 기법은 모델을 재탈옥하기 어렵게 만들지만, 불가능하게 하지는 못한다. 현재 대부분의 모델 탈옥 비용이 낮기 때문에, 실제 공격자들이 목표 달성을 위해 비용이 큰 공격을 사용하지 않는 이유를 설명할 수 있다.
문제 해결 사이클: 안전성과 보안에 대한 보장이 없는 상황에서는, 가능한 한 깨지기 어려운 AI 시스템을 개발할 방법이 필요하다. 한 가지 방법은 문제 해결 사이클을 사용하는 것으로, 여러 차례의 레드팀과 완화 조치를 반복하여 시스템이 다양한 공격에 견고하도록 만드는 것이다. 이 접근법을 Microsoft의 Phi-3 언어 모델을 안전하게 정렬하는 데 적용했으며, 다양한 피해와 시나리오를 다루었다. 완화 조치가 의도치 않게 새로운 위험을 초래할 수도 있기 때문에, 공격과 방어 전략을 지속적으로 적용하는 퍼플팀 방법이 단일 레드팀 사이클보다 공격 비용을 높이는 데 더 효과적일 수 있다.
정책 및 규제: 규제는 여러 방식으로 공격 비용을 높일 수 있다. 예를 들어, 규제는 조직이 엄격한 보안 관행을 준수하도록 요구하여 산업 전반의 방어 수준을 높일 수 있다. 또한, 법률은 불법 활동에 대한 명확한 결과를 규정함으로써 공격자를 억제할 수 있다. AI 개발과 사용을 규제하는 것은 복잡하며, 전 세계 정부는 혁신을 저해하지 않으면서 이러한 강력한 기술을 어떻게 통제할지 고민하고 있다. 설사 AI 시스템이 합의된 규칙을 준수하도록 보장할 수 있다 하더라도, 이러한 규칙은 우선순위 변화에 따라 시간이 지남에 따라 필연적으로 변할 것이다.

안전하고 보안이 강화된 AI 시스템을 구축하는 작업은 결코 완전히 끝나지 않는다. 그러나 공격 비용을 높임으로써, 오늘날의 프롬프트 주입이 결국 2000년대 초반의 버퍼 오버플로와 유사한 위치에 놓이게 될 것이라고 믿는다. 완전히 제거되지는 않겠지만, 다층 방어 조치와 안전 우선 설계를 통해 대부분 완화될 것이다.

3 열린 질문(Open questions)

지난 몇 년간 AI 레드팀 경험을 통해 얻은 교훈을 바탕으로, 몇 가지 열린 질문을 제시한다.
1) AI 레드팀은 새로운 능력과 등장하는 피해 영역에 맞춰 지속적으로 실무를 업데이트해야 한다. 특히, 설득, 속임수, 복제와 같은 LLM의 위험한 능력을 어떻게 탐지해야 할까? 또한, 영상 생성 모델에서 어떤 새로운 위험을 탐지해야 하며, 현존 최첨단 모델보다 발전된 모델에서 어떤 능력이 나타날 수 있을까?
2) 모델이 점점 다국어를 지원하고 전 세계에 배포됨에 따라, 기존 AI 레드팀 실무를 다양한 언어적·문화적 맥락으로 어떻게 번역할 수 있을까? 예를 들어, 여러 배경을 가진 사람들의 전문성을 활용하는 오픈소스 레드팀 이니셔티브를 시작할 수 있을까?
3) 조직이 그들의 방법과 결과를 명확하게 전달할 수 있도록 AI 레드팀 실무를 어떤 방식으로 표준화해야 할까? 본 문서에서 설명한 위협 모델 온톨로지가 올바른 방향으로의 한 걸음이라고 생각하지만, 개별 프레임워크는 종종 지나치게 제한적일 수 있음을 인식한다. 다른 AI 레

드팀이 제안한 온톨로지를 모듈식으로 다루고, 연구 결과를 요약, 추적, 전달하기 쉽게 만드는 추가 도구를 개발할 것을 권장한다.

4 결론(Conclusion)

AI 레드팀은 AI 시스템이 야기할 수 있는 안전 및 보안 위험을 식별하기 위한 초기 단계이자 빠르게 진화하는 실무이다. 전 세계의 기업, 연구기관, 정부가 AI 위험 평가를 수행하는 방법에 대해 고민하는 가운데, Microsoft에서 100개 이상의 생성형 AI 제품을 대상으로 수행한 레드팀 경험을 바탕으로 실질적인 권고사항을 제공한다. 내부 위협 모델 온톨로지, 8가지 주요 교훈, 5개의 사례 연구를 공유하며, 레드팀 활동을 실제로 발생할 가능성이 높은 피해와 정렬하는 방법에 중점을 둔다. 다른 연구자와 실무자가 이러한 교훈을 기반으로 발전시키고, 여기서 제시한 열린 질문들에 답해 나가기를 권장한다.